DOCUMENT RESUME

ED 263 155 TM 850 637

AUTHOR Wilson, Mark
TITLE Measuring Stages of Growth: A Psychometric Model of Hierarchical Development. Occasional Paper No. 19.
INSTITUTION Australian Council for Educational Research, Hawthorn.
REPORT NO ISBN-0-85563-397-2
PUB DATE Jan 85
NOTE 95p.
AVAILABLE FROM The Australian Council for Educational Research, Frederick Street, Hawthorn, Victoria 3122, Australia.
PUB TYPE Reports - Research/Technical (143) -- Books (010)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Cognitive Development; *Cognitive Measurement; *Developmental Stages; Elementary Education; Estimation (Mathematics); Foreign Countries; Latent Trait Theory; *Learning Theories; *Mathematical Models; Preschool Education; Subtraction; Vertical Organization; Young Adults
IDENTIFIERS Australia; Gagnes Taxonomy; *Hierarchical Learning; Piagetian Tasks; Piagetian Theory; *Saltus Model

ABSTRACT

A psychometric model called Saltus, which represents the qualitative aspects of hierarchical development in a form applicable to additive measurement, was applied. Both Piaget's theory of cognitive development and Gagne's theory of learning hierarchies were used to establish the common features of hierarchical development: (1) gappiness--the logical construction of the hierarchy which occurs when there is no state between adjacent stages; and (2) rigidity--learning behavior, exhibited by a fixed sequence of progression through stages. Saltus assumes a theory with gappiness expressed through items or tasks, and estimates the rigidity of data, thus testing the hypothesized gappiness. Four experiments tested the Saltus model in elementary school Australian children and with subjects aged 3, 4, 5, 8, 12, and 21 years. Three Piagetian tasks of rule assessment involving proportionality (balance scale, projection of shadows, and a probability task) gave clear evidence of rigidity in the step from the pre-operational stage to the concrete operational stage. The next step, to the formal operational stage, did not show rigidity, although gappiness was evident. The hypothesized existence of a gap to split the concrete operational stage was not supported. The Gagnean data on a subtraction task showed strong rigidity. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED263155

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Williams, Jeff

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

MEASURING STAGES OF GROWTH:
A Psychological Model of Hierarchical Development

Mark Wilson

Occasional Paper No. 19 - January 1985

Australian Council for Educational Research
Radford House, 9 Frederick Street,
Hawthorn, Victoria 3122, Australia

1985

Published by
The Australian Council for Educational Research,
Frederick Street, Hawthorn, Victoria 3122

Printed and bound by Brown Prior Anderson Pty Ltd,
Burwood, Victoria 3125

National Library of Australia Cataloguing-in-Publication data.

Wilson, Mark, 1954-
Measuring stages of growth.

Bibliography.

ISBN 0 85563 397 2.

1. Cognition in children - Mathematical models.
2. Psychometrics. I. Australian Council for Educational Research. II. Title. (Series: Occasional paper (Australian Council for Educational Research); no.19).

155.4'13

Copyright © ACER 1985

No part of this book may be reproduced in any form without permission from the publisher.

CONTENTS

		Page
LIST OF TABLES		vi
LIST OF FIGURES		viii
ACKNOWLEDGMENTS		xi
ABSTRACT		xiii
CHAPTER 1	THE CONCEPT OF A DEVELOPMENTAL HIERARCHY	1
	Introduction	
	The Theory of Piaget	2
	The Theory of Gagne	4
	A Generic Theory of Hierarchical Development	6
	The Rasch Model	7
	Application of the Rasch Model to Developmental Hierarchies	10
	An Adaptation of the Rasch Model	11
CHAPTER 2	THE SALTUS MODEL	15
	Introduction	
	Logit Scale Representation of the Saltus Matrix	16
	Relationship to the Rasch Model	21
	Relationship to the Generic Theory	23
	Estimation of Parameters	25
	Assessing the Fit of Data to the Model	28
	Checking the Performance of Saltus	29
CHAPTER 3	A SUBTRACTION HIERARCHY	30
	The Subtraction Tasks	
	Comparison with the Rasch Analysis	31
	Replicating the Saltus Results	41
CHAPTER 4	A RULE-ASSESSMENT HIERARCHY	45
	Introduction	
	The Balance Scale Task	50
	Linking the Saltus Analyses	64
CHAPTER 5	CONCLUSION	72
	Background to the Saltus Model	
	Description of the Saltus Model	
	Application of Saltus	74
	Implications of the Research	79
REFERENCES		81

LIST OF TABLES

		Page
Table 1.1	Correspondences Amongst the Theories	5
Table 2.1	The Saltus Matrix	16
Table 2.2	Probability of Success	17
Table 3.1	Subtraction Objectives	31
Table 3.2	Sample Used for Saltus Analyses	33
Table 3.3	Rasch Estimates for the 3V4 Sample	35
Table 3.4	Saltus Estimates for the 3V4 Sample	36
Table 3.5	Score Estimates for the Rasch and Saltus Analyses	38
Table 3.6	Probability of Success on Easiest and Hardest Items	
Table 3.7	Saltus Item Fit Statistics for the 3V4 Sample	37
Table 3.8	Saltus Student Fit Statistics from the 3V4 Sample	40
Table 3.9	Subtraction Saltus Matrices, with Standard Errors in Parentheses	42
Table 3.10	Subtraction Asymmetry Indices, with Standard Errors in Parentheses	
Table 3.11	Summary of Saltus Analyses for Victorian Boys and Girls	44
Table 4.1	Siegler Predictions for the Balance Scale Task	49
Table 4.2	Rasch Results for Balance Scale	51
Table 4.3	Number of Subjects in the Balance Scale Analyses	52
Table 4.4	Simulations for the E to S Step	
Table 4.5	Item Estimates for the E to S Step	53
Table 4.6	Score Estimates for the E to S Step	54
Table 4.7	Siegler and Saltus Classifications for the E to S Step	55
Table 4.8	Simulations for D to S Step	57
Table 4.9	Item Estimates for the S to CS Step	58
Table 4.10	Score Estimates for the S to CS Step	
Table 4.11	Siegler Predictions and Saltus Estimates of Success for the S to CS Step	59
Table 4.12	Saltus Item Estimates for the S to CE Step	62

		Page
Table 4.13	Score Estimates for the S to CE Step	62
Table 4.14	Siegler Predictions and Saltus Estimates for the S to CE Step	63
Table 4.15	Saltus Matrices from First and Second Testings of the Balance Scale Task	66
Table 4.16	Saltus Statistics for the Second Testing of the Shadow Task	
Table 4.17	Siegler Predictions for the Probability Task	67
Table 4.18	Saltus Statistics for the First Testing of the Probability Task	68
Table 4.19	Common Subjects Used to Link the Three Logit Scales	69

LIST OF FIGURES

		Page
Figure 1.1	An Example of a Rasch Scale	8
Figure 1.2	Two Stages Represented on a Rasch Scale	11
Figure 2.1	Saltus Matrix and Logit Scale for Case (i)	20
Figure 2.2	Saltus Matrix and Logit Scale for Case (ii)	
Figure 2.3	Saltus Matrix and Logit Scale for Case (iii)	
Figure 2.4	Saltus Matrix and Logit Scale for Case (iv)	22
Figure 2.5	Saltus Matrix and Logit Scale for Case (v)	
Figure 2.6	Saltus Matrix and Logit Scale for Case (vi)	
Figure 3.1	Item Difficulties on the RAPT Subtraction Scale	32
Figure 3.2	Items in the Subtraction Tests	34
Figure 3.3	Rasch Estimates for the 3V4 Sample	35
Figure 3.4	Saltus Estimates for the 3V4 Sample	36
Figure 3.5	Rasch and Saltus Difficulties for Type A Items	37
Figure 3.6	Rasch and Saltus Difficulties for Type B Items	
Figure 4.1	Siegler Rule I	46
Figure 4.2	Siegler Rule II	47
Figure 4.3	Siegler Rule III	48
Figure 4.4	Siegler Rule IV	49
Figure 4.5	Group I Gaps for the E to S Simulations	53
Figure 4.6	Logit Scale for the Balance Scale E to S Step	54
Figure 4.7	Siegler Compared to Saltus for the E to S Step	56
Figure 4.8	Group I Gaps for the D to S Step	57
Figure 4.9	Logit Scale for the Balance Scale S to CS Step	59
Figure 4.10	Siegler Compared to Saltus for the S to CS Step	60
Figure 4.11	Logit Scale for the Balance Scale S to CE Step	63
Figure 4.12	Siegler Compared to Saltus for the S to CE Step	64
Figure 4.13	Linked Logit Scale for the Balance Scale Task	65
Figure 4.14	Linked Logit Scale for the Shadows Task	67

		Page
Figure 4.15	Linked Logit Scale for the Probability Task	69
Figure 4.16	Linked Saltus Logit Scale for the Three Tasks	70
Figure 4.17	Linked Rasch Logit Scale for the Three Tasks	71
Figure 5.1	Rasch Estimates for 3-digit Subtraction Items	75
Figure 5.2	Saltus Estimates for 3-digit Subtraction Items	76
Figure 5.3	Siegler vs. Saltus for Balance Scale	78
Figure 5.4	Linked logit Scale for the Probability Task	79

ACKNOWLEDGMENTS

For encouraging me to come to Chicago, I would like to thank John Keeves and John Izard. I am grateful to Larry Hedges and Zalman Usiskin for many helpful suggestions and comments on this work, and also to Janellen Huttenlocher and Tom Trabbasso for their thoughtful reviews of the application of the Saltus model. For their generosity in allowing me to re-analyse their data, and for the care and attention to substantive theory that are displayed in the data, I am deeply indebted to John Izard, Stephen Farish and Graham Ward (the Subtraction data), and Robert Siegler (the Rule-assessment data). I am especially indebted to Ben Wright, the chairman of my committee, for providing the stimulus and encouragement for this work, for guiding it through its many changes, and for having the patience to wait for it to develop.

For providing the daily support and encouragement without which this work would not have been possible, I would like to thank Janet Williams.

ABSTRACT

Cognitive scientists have proposed many theories of intellectual development. Prominent among these have been theories that describe a child's growth through a sequence of hierarchical stages. Psychometricians too, have developed many models for the measurement of intellectual abilities. But there has been little contact between these two branches of scientific endeavour. The psychometric models which have been applied to developmental hierarchies have either not done justice to the complexity of the hierarchies, or have been inadequate in their assumptions about the measurement process.

This research has derived and applied a psychometric model, called Saltus, which represents the qualitative aspects of hierarchical development in a form which can lead to additive measurement.

Two theories of development - Piaget's theory of cognitive development and Gagné's theory of learning hierarchies - were used to establish the common features of hierarchical development. These are: gappiness, which pertains to the logical construction of the hierarchy and occurs when there is no state between adjacent stages, and rigidity, which pertains to the behaviour of learners, and is exhibited by a fixed sequence of progression through stages. Saltus assumes a theory with gappiness expressed through items or tasks and estimates the rigidity of the data, thus testing the hypothesized gappiness.

Four data sets, collected by researchers working within the traditions of Piaget and Gagné, were used to explore the usefulness of the Saltus model under practical application.

The three Piagetian data sets gave clear evidence of rigidity in the step from the pre-operational stage to the concrete operational stage. The next step, to the formal operational stage, did not show rigidity, although gappiness was evident: this was associated with item designs that elicited guessing and failed to produce homogeneous item difficulties. In addition, the existence of a gap, hypothesized by the experimenter to split the concrete operational stage, was not supported by the Saltus results. The Gagnean data - produced by constructed-response subtraction items that span the step to learn regrouping - showed strong rigidity. This rigidity was displayed, with only small variation, under changes in the stimuli (2-digit and 3-digit items), age of the students (Year 3 and Year 4) and geographical location (different Australian states).

CHAPTER 1

THE CONCEPT OF A DEVELOPMENTAL HIERARCHY

Introduction

The meaning of the word 'development' given by the Oxford English Dictionary (1961) is: 'The growth or unfolding of what is in the germ.' Its meaning for cognitive scientists, however, has filtered down through its interpretation in a number of other sciences. A principal one is biology in its embryological and evolutionist parts. Sir Ernest Nagel delineated this narrower meaning. According to him, what biologists mean by 'development' is:

> . . . a sequence of continuous changes eventuating in some outcome, however vaguely specified, which is somehow potentially present in the earlier stages of the process . . . (The) changes must be cumulative and irreversible . . . those changes must in addition eventuate in modes of organisation not previously manifested in the history of the developing system. (Nagel, 1957, pp.15-16)

Nagel's concept has been assumed into the cognitive sciences as the basic idea of psychological development, and will be used as the starting point for discussion of developmental hierarchies in this work. It includes references to 'stages' and 'modes of organization' which suggest qualitative changes between steps in a hierarchy.

There are many aspects to the development of a human being: here we are interested in development as learning. In the study of learning, there are three foci - the learner, the teacher, and the matter to be learned. Equivalently, there are three meanings which are commonly ascribed to the concept of a 'hierarchy' in development:

1 a psychological sequence, the order in which a topic can be learned by a child;

2 an instructional sequence, the order in which a topic is taught by the teacher;

3 a logical sequence, inherent within the topic to be learned, reflecting the basic structure of the topic.

To these I wish to add a fourth concept:

4 an empirical sequence, the order in which children are observed to learn a topic.

These four types of sequence are distinct, but in a given context are necessarily inter-related. An instructional sequence is available for scrutiny, through observation of teachers' behaviour: a logical sequence can be exposed by analysis of the concepts and skills used in a topic: an empirical sequence reveals itself in the test results or behaviour of the child. However, the psychological sequence occurs within the child, where it cannot be observed. This problem was of concern to Max Weber (1904-1949) who asserted that developmental sequences could be constructed into ideal types. He

described the relationship between such an ideal type and the course of development of a particular society thus:

> Whether the empirical-historical course of development was actually identical with the constructed one can be investigated by using this concept as a heuristic device for the comparison of the ideal type and the 'facts' . . . This procedure gives rise to no methodological doubts so long as we keep in mind that ideal-typical developmental constructs and history are to be sharply distinguished from each other, and that the construct here is no more than the means for explicitly and validly imputing an historical event to its real causes while eliminating those which on the basis of our present knowledge seem impossible. (Weber, 1904-1949, pp.101-102)

This problem of how to use the concept of developmental sequences has been carried over into cognitive science and lies behind the addition of the fourth type of hierarchical sequence - the psychological sequence corresponds to a Weberian 'ideal type' and the empirical sequence to his 'facts', whereas the other two have some features common to both.

The Theory of Piaget

The premier theory of cognitive development today must be that of Jean Piaget:[1] he was initially trained as a biologist and many of his concepts reveal this background. Piaget's theory is a theory of the development of structure in intelligent behaviour. He distinguished structure from the content of intelligent behaviour, which are the particulars of any situation, the environment, the stimuli, and the psychomotor abilities of the child. And he distinguished both structure and content from the function of intelligent behaviour, which are those aspects which hold constant across all situations, and is the means by which a child develops from one structure to the next. Function is a concept of biological origin, the main components of which can be summarized as:

> For Piaget, intelligence is not something which is qualitatively fixed at birth, but rather, is a form of adaptation characterised by equilibrium. Part of man's biological inheritance is a striving for equilibrium in mental processes as well as in physiological processes. Twin processes are involved: assimilation and accomodation. The child assimilates information from the environment which may upset existing equilibrium, and then accommodates present structures to the new so that equilibrium is restored. (Stendler, 1967, p.336)

This dynamic aspect of intelligence operates to move the child through a series of qualitatively distinct stages each characterized by a hierarchy of different structures.

---

1 The Piagetian literature is too voluminous to cite, so the interested reader can refer to Flavell's summary (Flavell, 1963) for general matters. Specific reference to Piaget's and his colleagues' work will be made only when the discussion is detailed.

Four stages are used to seriate intellectual development. The Sensorimotor Period lasting from 0 to 2 years is characterized by differentiation of self from others, the attainment of object permanence, the acquisition of manipulative skills to seek and maintain interesting stimuli, and a primitive understanding of causality, time and space. The Preoperational Period, lasting from about 2 to 6 years, is characterized by the development of symbolic functions in language and the dominance of irreversibility, centration and egocentricity in problem solving. The Concrete Operational Period, lasting from about 6 to 11 years, draws its name from the successful application of logical thinking based on reversibility, decentration and the ability to take the role of others, to concrete problems in the real world; the conservation of mass, weight and volume are developed during this period. The final stage defined is that of Formal Operations: here formal reasoning is applied to complex and possibility abstract problems.

These stages are more than descriptive tags for period of chronological age, or bundles of attributes which have been observed to cohere. Piaget used them as a theoretical tool with which to analyse behaviour, and so he needed to provide a sound theoretical definition for them. The criteria he provided (Piaget, 1960, pp.13-15) are:

1 A fixed order of succession: the age at which certain stages are attained may vary between individuals, but the stages must be attained in a fixed order by an individual. This was called hierarchization by Pinard and Laurendeau (1969).

2 Each stage must be subsumed into the next: for instance, the concrete problems mastered in the concrete operations period are integrated into understanding at the formal operations level as applications of general principles. This was called integration by Piaget.

3 Attainment of a stage must solve (logical) problems arising through the application of the structures of the previous stage and must lay the seeds for the apprehension of the problems which will be solved in the next stage. The first and last stages cannot, of course, fulfill both criteria. This was called consolidation by Pinard and Laurendeau (1969).

4 All the characteristics of a stage, all the preparations for it, and all the achievements possible within it, must form one general structure; appendages not systemically connected with the whole are not to be considered part of a stage. Structuring was the name given to this by Pinard and Laurendeau (1969).

5 Each stage must represent an equilibrium level, and the succession of stages should show a broadening of content and an increase in the stability of the equilibrium. This was called equilibriation by Piaget.

Two important observations were made by Piaget concerning his conception of stages (Piaget, 1960, p.16). First, he described a stage theory with just hierarchization as a minimum program, and one satisfying all five criteria as a maximum program. He was not dismissing stage theories which did not meet his standards, but was drawing attention to the theoretical drawbacks of such constructs. He noted, for example, that the stages in Freud's psycho-analytic theory do not exhibit integration, but found great merit in Erikson's stages because they do (Erikson, 1956). Second, a very general point, but one which should always be considered in the analysis of Piagetian ideas: Piaget's theory does not concern itself with the idiosyncracies of individuals. At a conference bringing together Piagetians and psychometricians, Piaget defined his attitude to

> . . . ordinal succession, not in general development, but in the development of the individual . . . This, I must confess, is a problem I have unfortunately never studied, because I have no interest whatsoever in the individual. I am very interested in general mechanisms, intelligence and cognitive functions, but what makes one individual different from another seems to be - and I am speaking personally and to my great regret - far less instructive as regards the study of the human mind in general. (Piaget & Inhelder, 1971, p.211)

Thus, to Piaget, the idea of a comprehensive theory, that explained the behaviour of every individual under every circumstance, was nothing but a red herring.

The Piagetian theory is a psychological theory of hierarchical development. It is not necessarily a theory of instruction, although it provides definite limitations on the potentialities of instruction (Stendler, 1967, p.343). There are many elements in the details of his theory, mainly in the structuring concepts such as groupings and groups which owe much to modern abstract algebra. Although the model of the final goal, the formal operations stage, is derived from the logic of the situation, the stages leading to the acquisition of that final stage do not represent adult 'logic'. The 'logical' analysis of a problem will not necessarily reveal the Piagetian stages through which a child would need to pass in order to master it.

The Theory of Gagné

R.M. Gagné, working in the testing and training of servicemen during World War II, concentrated not on the psychological state of the learner, nor on the structure of an area of knowledge, but on an analysis of the task to be taught (Gagné, 1962b). His technique consists of:

1 identifying a pinnacle skill to be mastered, and
2 establishing a set of subordinate skills by successively laying out the prerequisites for each skill.

These subskills then form a hierarchy if:

Table 1.1 Correspondences Amongst the Theories

Piaget	Gagné
Hierarchization	Rigidity
Integration	-
Consolidation	-
Structuring	Gappiness
Equilibrium	-

(a) no individual could perform the final task without having these subordinate capabilities . . . and

(b) that any subordinate task in the hierarchy could be performed by an individual provided suitable instructions were given, and provided the relevant subordinate knowledges could be recalled by him. (Gagné, 1962a, p.356)

Criterion (a), which I shall call rigidity, is similar to Piaget's hierarchization. The meaning of the second criterion, however depends on one's concept of 'suitable' and 'relevant'. This criterion could be interpreted as a proxy for integration and adjacency. The tasks contained in these hierarchies are not unrestricted, however. They must represent

> . . . the kind of change in human behavior which permits the individual to perform successfully on an entire class of specific tasks, rather than simply on one member of the class . . . (Gagné, 1962a, p.355)

This restriction was later used to exclude verbal knowledge (Gagné, 1968), but in its wider interpretation it seems equivalent to Piaget's structuring: I shall call this gappiness (See Table 1.1).

Gagné's theory can be interpreted as a theory of instruction, founded on an analysis of the skills to be mastered, but having as an essential requirement that the analysis must result in subskills which can be taught successively to the student. Unlike Piaget, Gagné does hold that this is genuinely a theory of individual behaviour, and so for a postulated hierarchy to be accepted, it must be shown, within experimental error, to hold for every individual investigated. Several early studies (Gagné & Bassler, 1963; Gagné, Mager, Garstons & Paradise, 1962; Gagné & Paradise, 1961) revealed that while most of the learners did behave consistently according to the postulated learning hierarchies, some did not. This disappointment spawned a series of increasingly sophisticated statistical procedures for the "validation" of learning hierarchies. Gagné's effort to explain the idiosyncracies of individuals has not met with success and workers in this field have since lowered their expectations to that of finding 'reasonably accurate hierarchies' (White, 1981, p.227).

Cognitive scientists have produced these complicated theories of development because it is the qualitative rather than the quantitative aspects of cognitive processing that are more interesting. Cognitive scientists are more interested in changes in the organisation of thought than in such quantitative matters as the number of concepts that a child may possess (Flavell & Wohlwill, 1969, p.77). It is the aim of this work to develop and apply a model that incorporates some of the qualitative aspects of these theories. This model brings the three types of learning sequences - psychological, instructional and logical - together at the common level of the empirical behaviour of learners, that is, at the measurement level.

The first requirement for constructing a model appropriate for the measurement of developmental hierarchies is a clear picture of what constitutes such a hierarchy. The previous sections have described versions of developmental hierarchies and discussed some problems with them. In order to proceed, however, it is necessary to construct a generic theory of hierarchical development, concentrating the essential elements of each of the individual theories described above.

The properties of this generic theory of hierarchical development are gappiness and rigidity. A hierarchy exhibits gappiness when, according to the theory, there is no possible state between stages. Such a gap is represented in Piaget's theory by 'structuring', and in Gagné's theory by the restriction which I named 'gappiness'.

A hierarchy exhibits rigidity when, according to the theory, a child at a particular stage of the hierarchy must have passed through each stage below. This is equivalent to the hierarchization of Piaget, and the first criterion used by Gagné which I called rigidity. The other properties have been left out because they are idiosyncratic to the theory in which they occur. The exception is integration-adjacency, which pertains to the substantive meaning of the stages in the hierarchy rather than their structure, and thus, cannot display itself directly. Since integration-adjacency is responsible for rigidity, this property is contained in the generic model, through its consequences.

These two features - rigidity and gappiness - are together the defining elements of the generic theory of developmental hierarchies which will be examined in this work. How can they be embodied in a psychometric model for the analysis of data? The psychometric model must work at the level of measurement, and is, therefore, subject to the problems of connecting qualitative (i.e. theoretical) concepts with quantitative events. Thus a third element, that of the uncertainty of data, must be incorporated when attempting to bring the generic theory of hierarchical development to life through a psychometric model.

Theories about cognitive development are not well accepted in the scientific community without adequate empirical demonstration of their major predictions. The

analysis of the data generated by experiments to test and apply all but the most simple-minded of such theories demands appropriate measurement models. In the case of hierarchical theories of development, the special features of these theories, gappiness and rigidity, impose particular requirements on any measurement model used to analyse the data produced by applications and testings of the theories.

The Rasch Model

The Rasch model (Rasch, 1960/1980) for the analysis of psychometric data is a way to place persons and items on a scale with a clear probabilistic interpretation of distance on the scale. For a dichotomously-scored item j, with difficulty δ_j, attempted by person i, with ability β_i, the probability of a correct response, $y_{ij} = 1$, is modelled as (Wright and Stone, 1979, p.15)

$$P(y_{ij}=1) = \frac{\exp(\beta_i - \delta_j)}{1 + \exp(\beta_i - \delta_j)} \qquad (1)$$

$$= \Psi(\lambda_{ij})$$

where $\lambda_{ij} = \beta_i - \delta_j$,

and Ψ is the logistic function defined by

$\Psi(x) = \exp(x)/(1+\exp(x))$.

The probability of an incorrect response is modelled as

$$P(y_{ij}=0) = \frac{1}{1 + \exp(\beta_i - \delta_j)} .$$

When combining the probabilities of L items to find the probability of a response vector $\underline{y}_i = (y_{i1}, \ldots, y_{iL})$, local independence is assumed. That is, with a vector of item difficulties, $\underline{\delta} = (\delta_1, \ldots, \delta_L)$ the probability of response vector $\underline{y}_i$ is

$$P(\underline{y}_i \mid \beta_i, \underline{\delta}) = \prod_{j=1}^{L} P(y_{ij} \mid \beta_i, \underline{\delta}) .$$

Local independence says that in calculating the probability of each response, one must take into account the ability of the person and the difficulty of the item, but once that has been done a simple multiplicative rule tells how to combine the probabilities.

Using local independence, the probability of person i scoring t on the set of L items is

$$P(\sum_{j=1}^{L} y_{ij} = t \mid \beta_i, \underline{\delta}) = \sum_{\underline{T}} P(\underline{y}_i \mid \beta_i, \underline{\delta})$$

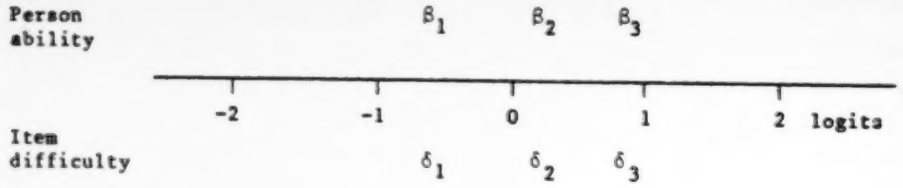


Figure 1.1 An Example of a Rasch Scale

where T is the set of all $\underline{y}_i$ with $\sum_{j=1}^{L} y_{ij} = t$.

This assumption of local independence would be incorrect if some subgroup of persons has a special relationship with some subgroup of items that is not encompassed by the relationship.

$$\lambda_{ij} = \beta_i - \delta_j .$$

This simple relationship between the person and item parameters allows a straightforward representation of the Rasch scale. Consider three persons with abilities β_1=-1, β_2=0, β_3=1, and three items with difficulties δ_1=-1, δ_2=0, δ_3=1 as represented in Figure 1.1. Person 2 has ability equal to item 2, so, using equation (1), he has a 50 per cent chance of getting the item right:

$$P(y_{22}=1) = \frac{\exp(0-0)}{1+\exp(0-0)}$$

$$= \frac{1}{1+1}$$

$$= 0.5 .$$

The probability of person 2 getting item 3 right is

$$P(y_{23}=1) = \frac{\exp(0-1)}{1+\exp(0-1)}$$

$$= 0.27 .$$

The scale is equal interval, that is, the (signed) distance between the person ability and the item difficulty governs the probability of a correct response, and the distance has the same meaning no matter where on the scale it is located. Thus, the probability of a correct response from person 1 on item 2 is

$$P(y_{12}=1) = \frac{\exp(-1-0)}{1+\exp(-1-0)}$$

$$= 0.27$$

which is the same as the probability of a correct response from person 2 on item 3: the location of person and item changed, but the distance between them did not.

The natural unit for the scale is the 'logit': this is interpreted as the (natural) logarithmic odds of success: if P is the probability of success corresponding to one logit, then

$$\log(P/(1-P)) = 1$$

so

$$P = \frac{\exp(1)}{1 + \exp(1)}$$

$$= 0.73 .$$

Thus a positive difference of one logit means a log odds of success of 1 and a probability of success of 0.73. The origin of the scale is arbitrary; it is usually chosen as the average of the items because the items are explicit and so form a more interpretable reference point than would be obtainable from the people.

The Rasch model allows separable estimation of the parameters, that is each person and item parameter and its associated statistics can be expressed as a separate multiplicative component of the modelled likelihood of the data (Rasch, 1960/1980, pp.171-172).

Conditional probabilities can eliminate the person parameters from item calibration. Item parameters can be eliminated in the estimation of person parameters in the same way. This is what is meant when the Rasch parameters are said to be 'test-free' and 'sample-free' (Wright and Douglas, 1977).

This attribute of the Rasch model is equivalent to the existence of simple sufficient statistics for both persons and items. A sufficient statistic for a parameter is one that contains all the modelled information in the data regarding that parameter; in this sense a sufficient statistic is a 'best' statistic. A statistic t estimating a parameter β over a sample $\underline{x}$, is sufficient if the likelihood L can be expressed as a product of two functions, L_1 and L_2, the first of which involves the parameter and the statistic, and the second of which does not involve the parameter (Kendall and Stuart, 1969, p.9):

$$L(\beta, \underline{x}) = L_1(t, \beta)L_2(\underline{x}) . \qquad (2)$$

For the Rasch model, considering a person of ability β_i, the likelihood L, given a vector of item difficulties $\underline{\delta}$, is

$$L(\beta_i, y_i \mid \underline{\delta}) = \frac{\exp \sum_j y_{ij}(\beta_i - \delta_j)}{C(\beta_i, \delta)}$$

where $C(\beta_i, \underline{\delta}) = \prod_j (1 + \exp(\beta_i - \delta_j))$

so

$$L(\beta_i, y_i \mid \underline{\delta}) = \frac{\exp \beta_i(\sum_j y_{ij})}{C(\beta_i, \underline{\delta})} \exp(\sum_j y_{ij}) .$$

Thus the likelihood can be expressed in the required form and

$$r_i = \sum_j y_{ij},$$

the score for the person, is the required sufficient statistic for the person ability, given the item difficulties $\underline{\delta}$. Similarly,

$$s_j = \sum_i y_{ij}$$

is sufficient for δ_j given the person abilities $\underline{\beta}$. This is known as conditional sufficiency.

There are several estimation algorithms for finding the person and item parameters for a given data set. A good first approximation is given by PROX (Cohen, 1979; Wright and Stone, 1979, pp.28-45; Wright and Masters, 1982, pp.61-67), which assumes normal distributions of persons and items. The statistically best procedure is the algorithm CON (Andersen, 1972; Wright and Masters, 1982, pp.85-86), which calculates the symmetric functions necessary to achieve the conditional solution. For larger numbers of persons and items, however, this algorithm is cumbersome. An easier solution, UCON (Wright and Panchepaskan, 1969; Wright and Stone, 1979, pp.62-65; Wright and Masters, 1982, pp.72-80), is widely used for its simplicity, speed and accuracy: it is an iterative maximum likelihood method which estimates person and item parameters simultaneously.

Application of the Rasch Model to Developmental Hierarchies

When data from a test designed to identify a developmental hierarchy is analysed with the Rasch model, the resulting scale would be expected to exhibit segmentation. That is,

1 items representing different stages of the theory are contained in separate segments of the scale, with a non-zero distance between segments, and

2 segments are in the order predicted by the theory.

This definition is made in terms of parameters; if item estimates are being considered, then the idea of distance between the stages must include the standard error of measurement. Thus, for estimates, a non-zero distance between segments would be established by a difference between the closest items of adjacent segments of two or three times the standard errors of their calibrations. A useful indicator of segmentation is the segmentation index, S:

$$S = \delta_{Bmin} - \delta_{Amax},$$

on a scale where δ_{Bmin} is the difficulty of the easiest item of type B, and δ_{Amax} is the difficulty of the hardest item of type A.

Segmentation is the expression in Rasch terms, of the concepts of rigidity and gappiness. The gappiness of the stages defines the division of items into separate types

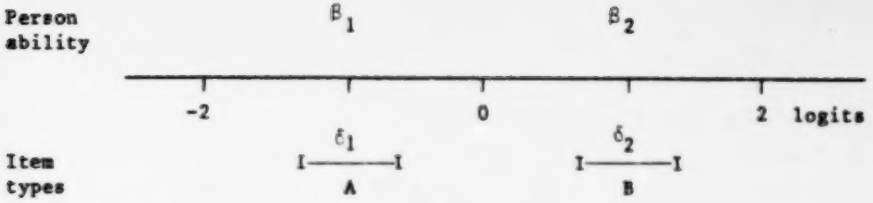


Figure 1.2 Two Stages Represented on a Rasch Scale

and the lack of intermediate states between stages means that distances between segments can be interpreted as indicating genuine stage gaps rather than an artifact of item selection. If overlap were found between sets of items representing adjacent stages, then some items of the lower stage must be more difficult than some items of the higher stage, which is inconsistent with the rigidity of a hierarchical theory.

Segmentation of the scale, however, is not completely analogous to rigidity. Consider, for example, the situation depicted in Figure 1.2 where two items, with difficulties δ_1 and δ_2, have been chosen to represent stages A and B, respectively. A person, with ability β_1, working through stage A has, according to the model, the following probability of success on item 2:

$$P(y_{12}=1) = \Psi(\beta_1 - \delta_2) = \Psi(-2) = 0.12 .$$

A person, with ability β_2, working through stage B has the same probability of getting item 1 wrong:

$$P(y_{21}=0) = 1 - \Psi(\beta_2 - \delta_1) = 1 - \Psi(2) = 0.12 .$$

The equality of $P(y_{21}=1)$ and $P(y_{12}=0)$ is caused by symmetry of the Rasch model: the model depends only on the logit distance between the stages. The Rasch estimation process determines this distance so that the two probabilities ($P(y_{12}=1)$ and $P(y_{21}=0)$ are equal, and the symmetry of the model is maintained. Compare this, however, with the concept of rigidity: persons must pass through the stages in a fixed order. Theoretically, a person passing through stage A cannot succeed on item type B, although some measurement error will inevitably occur: a person passing through stage B would be expected to do quite well on items of type A, but might not get them all correct, being subject to the same human error. There is nothing in this description requireing that the two sorts of error be equal. In general, we would expect $P(y_{12}=1)$ not to equal $P(y_{21}=0)$, and if care is taken to eliminate guessing on the items, we would expect

$$P(y_{21}=0) > P(y_{12}=1) .$$

The point is not that

$$P(y_{21}=0) = P(y_{12}=1)$$

could never occur, but that restricting the model so that it must occur does not express the theory of developmental hierarchies as embodied in the asymmetry of rigidity.

The fit of the Rasch model, however, can be used to detect rigidity. According to the discussion above, we expect

$$P(y_{21}=0) > P(y_{12}=1) . \qquad (3)$$

These probabilities can be expressed in terms of expected values calculated using the parameters

$$\pi_{12} = P(y_{12}=1) \quad \text{and}$$

$$1 - \pi_{21} = P(y_{21}=0) .$$

So equation (3) can be re-expressed as

$$1 - \pi_{21} > \pi_{12} . \qquad (4)$$

The Rasch model, however, will estimate the item difficulties and person abilities so that

$$1 - \pi_{21} = \hat{\pi}_{12} . \qquad (5)$$

where $\hat{\pi}_{21}$ and $\hat{\pi}_{12}$ are the expected values calculated using the Rasch estimates of the parameters. Suppose now that we accept the Rasch estimate of π_{21}, that is

$$\pi_{21} = \hat{\pi}_{21} .$$

Then equation (4) becomes

$$1 - \hat{\pi}_{21} > \pi_{12}$$

so that, using equation (5)

$$\hat{\pi}_{12} > \pi_{12} . \qquad (6)$$

This means that the estimated Rasch parameters will predict more success for persons in group I on items of type B than will actually occur.

With this in mind, the results of a Rasch analysis of items constructed according to a developmental hierarchy can be examined for symptoms of rigidity. The most obvious application of the above reasoning is to compare observed responses with those expected from the estimates. For persons in group I, the observed successes on type B items should be lower than those expected. This exercise is good for getting a 'feel' for the effect of a gap on person responses, but the detail becomes overwhelming as the number of persons increases. The solution is to consider the Rasch model item fit statistics.

These are calculated by summing the residuals over persons for each item and transforming the sums to a distribution close to a standard normal (Wright and Masters, 1982, pp.99-105). The theoretical residual is

$$x_{12} = y_{12} - \pi_{12} .$$

But estimates must be used to find the expected response, so instead we have an observed residual

$$\hat{x}_{12} = y_{12} - \hat{\pi}_{12} .$$

Which by equation (8) gives

$$x_{12} > \hat{x}_{12} .$$

Under the standardizing transformation, this discrepancy produces negative misfit values. Thus a pattern of negative fit statistics for items in stage B is a symptom of rigidity.

In a Rasch analysis we can expand our attention to more than two stages. Introducing a stage C above item type B, however, makes the situation more complicated, as one is unsure whether to view items of type B as an upper stage compared with type A or a lower stage compared with type C. It seems reasonable to think of the items in the lower portion of type B as the upper stage for type A items, and the items in the upper portion of type B as being the lower stage for type C items. Thus, in an analysis of several stages, one would look for a pattern of negative misfit at the lower end of successive stages.

This discussion has deduced expected patterns of misfit based on the assumption that the Rasch model is performing reasonably well in estimating the performances of person group II on item type A, but not so well in estimating the performance of person group I on items of type B. If the measurement situation led to a belief that other assumptions were more realistic, then different patterns of misfit would be expected and could be deduced in the same way as those above.

An Adaptation of the Rasch Model

The Linear Logistic Test Model (LLTM) (Fischer, 1973) is a Rasch model with a linear marginal condition developed to help explore the cognitive structures represented by the items in a Rasch scale. The form of the model is the same as that given above, with the item difficulty decomposed into a linear function of the weight and difficulty of the different cognitive operations which are assumed to be necessary for the successful completion of the item. The first step in applying the model is to assume a set of operations underlying the model. A weight describing the influence of each operation on

the item is then assigned: this is usually the number of times that the operation must occur in order to complete the item (Fischer, 1977, p.204). If there are m cognitive operations involved, the item difficulty is decomposed into:

$$\delta_j = \sum_{k=1}^{m} q_{jk} \eta_k + c$$

where q_{jk} is the weight of the operation k in item j,

η_k is the parameter attached to operation k,

and c is a normalizing constant.

The matrix (q_{jk}) must be of rank m for the parameters to be estimable. Estimation equations for the model were derived by Fischer (1972, 1973). The η_k parameters for the hypothesized cognitive operations are interpreted as the difficulties of the operations, although whether this is conceptualized as the difficulty of learning the operations or of performing the operations depends on the experimental situation (Spada, 1977, pp.243-249). The relevance of the results of an analysis using this model depends heavily on the plausibility of the weights assigned to the cognitive operations and one's certainty that the list is exhaustive (Spada and Kluwe, 1980, pp.29).

The application of this model to developmental hierarchies is subject to the same criticism as given in the preceding section for the application of the Rasch model to developmental hierarchies. Although the item parameters found by an LLTM analysis will not, in general, be the same as those found by a Rasch analysis, the same fundamental symmetry is present. The contribution that this model has made to the development of Saltus is the demonstration that the parameterization of the difficulties and abilities within the logistic function can be adapted to take into account certain special features of the measurement situation. Such an adaptation is the substance of the remainder of this work.

CHAPTER 2

THE SALTUS MODEL

Introduction

The word 'saltus' comes from the Latin for 'leap': the Oxford English Dictionary (1961) gives its meaning as 'a leap or sudden transition; a breach of continuity'. It has been chosen by this author as the name for this psychometric model because it embodies the twin notions of movement in a particular direction (i.e. rigidity) and jumpiness (i.e. gappiness) by which the generic theory of hierarchical development has been defined.

The interaction between a person i and an item j, recorded dichotomously as $y_{ij} = 1$ for correct and $y_{ij} = 0$ for incorrect, is governed by a logistic model:

$$P(y_{ij}=1) = \frac{\exp \lambda_{ij}}{1 + \exp \lambda_{ij}}$$

$$= \Psi(\lambda_{ij})$$

where the parameter λ_{ij} is composed of additive elements for person, β_i, item, δ_j, and also Saltus parameter, γ_{ij}:

$$\lambda_{ij} = \beta_i - \delta_j + \gamma_{ij}.$$

The Saltus parameter is not considered to vary by person and item, but by person group, h, and item type, k. Thus:

$$\gamma_{ij} = \gamma_{h(i)k(j)}$$

where h(i) is the group which contains person i,

and k(j) is the type of item j.

The groups and types are determined by the substantive theory. Item types are composed of items which, according to theory, represent particular stages. Person groups are then formed on the assumption that persons at or passing through a particular stage will score above the previous stage but not above the stage they are in. Thus, if there are L_A items of type A and L_B items of type B, persons scoring L_A or less are classified into group I and the remainder into group II. This is the basis for the classification used in the applications considered here: it has been chosen because it represents the expected pattern of responses that would occur if the hierarchical theory of development under consideration were correct. With this classification, the first person group is seen to be operating at the level of the first item type and the second person group is seen to be operating at the level of the second item type. Other ways of

Table 2.1 The Saltus Matrix

	Person	Group
Item type	I	II
A	γ_{AI}	γ_{AII}
B	γ_{BI}	γ_{BII}

using scores to classify persons are possible, but the one outlined above has given the clearest interpretation of the Saltus parameters, is consistent with the generic theory and provides an unambiguous first assessment of those who have not yet crossed the gap.

Were an external criterion superior to test score available, then this would be used to pre-classify the people. In that case, Saltus would not be needed to define the hierarchy, but could be used to co-ordinate a definition with other measures such as pencil and paper tests, multiple-choice tests, etc. Unfortunately, external criteria for developmental hierarchies are not available, nor perhaps, will ever be.

In order to make the presentation clearer, attention shall be restricted to cases where just two person groups and two item types are present. The core of the model, and the estimation algorithm in particular, do not need this restriction, but many of the interpretations become confusing if more stages are considered at one time.

With two person and two item groups, the Saltus parameters can be expressed in the Saltus matrix in Table 2.1 The arrangement given in the Table - item types indexed as rows A and B, and person types indexed as columns I and II - will be adhered to throughout.

Under the Saltus model, the probability of a correct response, $y_{ij}=1$, for person i in person group h, attempting item j of type k, is:

$$P(y_{ij}=1) = \frac{\exp(\beta_i - \delta_j + \gamma_{hk})}{1 + \exp(\beta_i - \delta_j + \gamma_{hk})} \quad . \qquad (7)$$

Note that, as person and item parameters occur in equation (7) combined with a Saltus parameter, interpretation of the person and item parameters must be made relative to the appropriate Saltus parameter. Probabilities for sets of items and people are combined using the assumption of local independence.

Logit Scale Representation of the Saltus Matrix

The Saltus parameters can be interpreted only in conjunction, with person and item parameters. For example, given a person of ability 0 and an item of difficulty 0 on the logit scale, the probability of that person getting the item correct is:

Table 2.2 Probability of Success

	Person	Group
Item type	I	II
A	$\Psi(\gamma_{AI})$	$\Psi(\gamma_{AII})$
B	$\Psi(\gamma_{BI})$	$\Psi(\gamma_{BII})$

$$\Psi(\gamma_{hk}) = \frac{\exp \gamma_{hk}}{1 + \exp \gamma_{hk}} \quad .$$

But the same person attempting an item of difficulty -1 logits would have probability $\Psi(1+\gamma_{hk})$ of succeeding. In order to simplify the discussion, we will suppose that person abilities do not vary within person groups and item difficulties do not vary within item types. Later, these restrictions will be eased so that only the average of the abilities within each group and the difficulties within each type need be 0. This focusses our attention on the Saltus parameters and the hierarchical step which they measure. Then the probabilities of success for the different person groups and item types are as given in Table 2.2

In order to represent this situation on a logit scale, we must set out the two item types and two corresponding person groups. Mark the location of item type A by d_A, item type B by d_B, person type I by b_I, and person type II by b_{II}. Then the Saltus matrix tells us what relationships among these locations to expect: the probability of a person in group I succeeding on an item of type A is, by equation (7).

$$P(y_{AI}=1) = \Psi(b_I - d_A),$$

but from Table 4,

$$P(y_{AI}=1) = \Psi(\gamma_{AI}),$$

hence,

$$b_I - d_A = \gamma_{AI} \quad .$$

Similarly,

$$b_I - d_B = \gamma_{BI}$$
$$b_{II} - d_A = \gamma_{AII}$$
$$b_{II} - d_B = \gamma_{BII}$$

This can be summarized in the matrix equation:

$$\begin{bmatrix} \gamma_{AI} & \gamma_{AII} \\ \gamma_{BI} & \gamma_{BII} \end{bmatrix} = \begin{bmatrix} b_I - d_A & b_{II} - d_A \\ b_I - d_B & b_{II} - d_B \end{bmatrix}$$

which shows what the Saltus parameters mean on the logit scale. This system of difference equations cannot be solved without a constraint. The location of item type A has been chosen as the reference point because, while persons in group II can be expected to have some reasonable failure rate on items of type A, the success rate of persons in group I on items of type B is expected to be irregular.

Setting

$$d_A = 0$$

the matrix equation becomes

$$\begin{bmatrix} \gamma_{AI} & \gamma_{AII} \\ \gamma_{BI} & \gamma_{BII} \end{bmatrix} = \begin{bmatrix} b_I & b_{II} \\ b_I - d_B & b_{II} - d_B \end{bmatrix} .$$

This gives,

$$b_I = \gamma_{AI} \quad \text{and}$$

$$b_{II} = \gamma_{AII} .$$

But two equations for d_B:

$$d_{BI} = b_I - \gamma_{BI} = \gamma_{AI} - \gamma_{BI} \quad \text{and}$$

$$d_{BII} = b_{II} - \gamma_{BII} = \gamma_{AII} - \gamma_{BII} .$$

Note that the solutions of these two equations have been denoted d_{BI} and d_{BII}. They will be called the group I gap and the group II gap, respectively. The difference between these two gaps is called the asymmetry index:

$$D = d_{BI} - d_{BII} = (\gamma_{AI} - \gamma_{BI}) - (\gamma_{AII} - \gamma_{BII}) .$$

When the gaps are equal, there is a unique placement for the items of type B and the asymmetry index is zero. If D is non-zero, then the items of type B cannot be given a unique location; the symbols BI and BII will then be used to refer to the different locations of item type B from the differing perspectives of person groups I and II, respectively. A positive asymmetry index indicates that the item types are relatively closer together (in terms of difficulty) for group II than for group I. This is consistent with the progression of difficulties for a developmental hierarchy; type B is almost impossibly harder than type A for persons in group I, but once the step between the stages has been straddled (that is, for persons in group II), the difference between the two item types becomes much less. Thus, when the asymmetry index is not zero in the examples and discussion that follow, it will be assumed that it is positive. A negative

asymmetry index indicates that the types are relatively closer together for group I than for group II. If this occurs, Saltus will estimate it; however, in a hierarchical situation, with segmentation between the item types, a negative asymmetry index is evidence of guessing or some similar fault in the item design. Thus negative asymmetry indices will not be discussed until they occur in Chapter 4.

The locations of item types A and B used in the definitions of the gaps and the asymmetry index, are, in the most general case, the mean locations of the items of types A and B. In the discussion in this chapter, the mean location of an item type is the same as the location of every item of that type since we have assumed that there is no variation within item types.

The following special cases will serve as signposts toward understanding the relationships between the Saltus parameters and the relationships between the Saltus parameters and the interaction of persons with items. In the interests of simplicity, the restriction that all person abilities are constant within groups and all item difficulties are constant within types will be maintained, although the interpretation of the diagrams is essentially the same with the lighter restriction that the average within each group and type is set of zero.

Case (i): Figure 2.1 Here the asymmetry index is zero and the segmentation index is also zero. The person groups and item types have no effect on person abilities and item difficulties. There is no segmentation of the item types and therefore there is no evidence of gappiness. Note that each Saltus parameter is named in the Saltus matrix.

Case (ii): Figure 2.2 Now the person groups are behaving differently, the first group sees the items as more difficult than the second group. Neither person group, however, has differentiated between item types. Again, the segmentation index is zero and the asymmetry index is zero (i.e. (c-c)-(d-d)=0).

Case (iii): Figure 2.3 Here the two person groups have different abilities, and the two item groups have different difficulties, but the person groups see the difference between the two item groups as equal. The segmentation index is a-b: the difference between the easiest item of type B, which is located at a-b, and the hardest item of type A, which is located at 0, is a-b, so there is some evidence of gappiness.

The asymmetry index is 0, that is

$$((c+a)-(c+b)) - ((d+a)-(d+b)) = (a-b) - (a-b) = 0$$

so there is no evidence of rigidity.

Case (iv): Figure 2.4 Case (iv) is a simplification of Case (iii). The asymmetry index is zero here also. This case will be used for simulations in which the person groups are located at their respective item types.

Saltus matrix

$$\begin{bmatrix} a\ (AI) & a\ (AII) \\ a\ (BI) & a\ (BII) \end{bmatrix}$$

Logit scale:

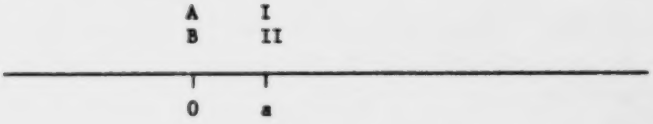


Figure 2.1 Saltus Matrix and Logit Scale for Case (i)

Saltus matrix

$$\begin{bmatrix} c\ (AI) & d\ (AII) \\ c\ (BI) & d\ (BII) \end{bmatrix}$$

Logit scale:

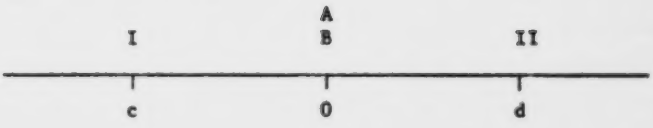


Figure 2.2 Saltus Matrix and Logit Scale for Case (ii)

Saltus matrix

$$\begin{bmatrix} c+a\ (AI) & d+a\ (AII) \\ c+a\ (BI) & d+a\ (BII) \end{bmatrix}$$

Logit scale:

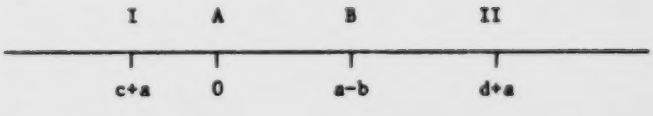


Figure 2.3 Saltus Matrix and Logit Scale for Case (iii)

Case (v): Figure 2.5 This is the most general expression for a Saltus matrix: there are no special relationships amongst the Saltus parameters. In order to clarify the presentation, the logit scales are presented first from the point of view of person group I, then person group II, then the two together. The asymmetry index is

$$D = (t-s) - (r-v)$$

which will, in general, not be zero. (The presentation in Figure 2.5 assumes that the asymmetry index is positive.) The group I segmentation index is t-s (the difference between BI and A), and the group II segmentation index is r-v (the difference between BII and A). The Figure exhibits both positive asymmetry and segmentation for both person groups.

Case (vi): Figure 2.6 This is specialization of Case (v). It will be used for simulations in which the person groups are located at their respective item types.

Relationship to the Rasch Model

The Saltus model preserves the basic features of the Rasch model while adding one other. This makes for a more complicated mode of presentation, however, and the advantages will have to be considered in each application. Under what conditions are the Saltus model and the Rasch model the same? For a Saltus matrix of the form

$$\begin{bmatrix} \gamma_{AI} & \gamma_{AII} \\ \gamma_{BI} & \gamma_{BII} \end{bmatrix}$$

the requirement for it to represent a Rasch model is that, using translations, we can apportion the Saltus parameters among the person and item parameters so that the Saltus matrix becomes null and the person and item parameters remain unique. This is what was attempted in the previous section when the logit scales for each person group were mapped onto one scale. This could be accomplished with a unique assignment of the person and item parameters when the asymmetry index was zero, and not otherwise. Thus a Saltus model is a Rasch model when the Saltus matrix has an asymmetry index of zero. When the asymmetry index is positive the Saltus model is estimating features in the data that the Rasch analysis can represent only as misfit. The further the asymmetry index is from zero, the less Rasch-like is the model.

Even when a Saltus model cannot be represented as a Rasch model, many of the features of the Rasch model persist. The event which occurs when an item is attempted by a person is now governed by a Saltus parameter as well as person and item parameters. It can still be represented on a logit scale with the added complication that the second type of item will have two locations depending on the group to which the person belongs. The location of item type B from the point of view of person group I,

Saltus matrix

$$\begin{bmatrix} 0\ (AI) & p\ (AII) \\ -p\ (BI) & 0\ (BII) \end{bmatrix}$$

Logit scale:

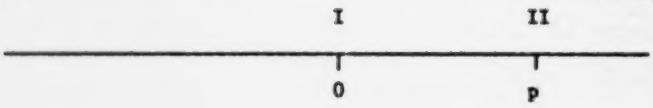


Figure 2.4 Saltus Matrix and Logit Scale for Case (iv)

Saltus matrix

$$\begin{bmatrix} t\ (AI) & r\ (AII) \\ s\ (BI) & v\ (BII) \end{bmatrix}$$

Logit scale for Person Group I:

Logit scale for Person Group II:

Combined logit scale:

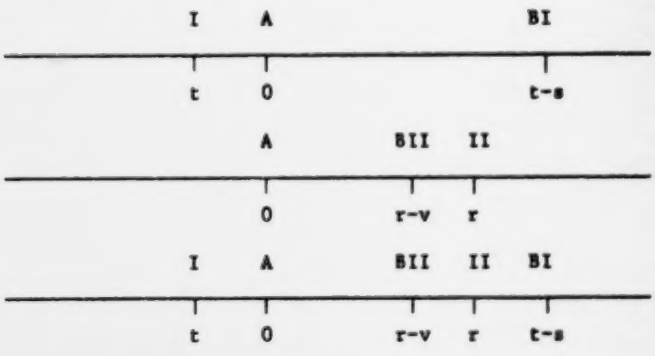


Figure 2.5 Saltus Matrix and Logit Scale for Case (v)

Saltus matrix

$$\begin{bmatrix} 0\ (AI) & u\ (AII) \\ -w\ (BI) & 0\ (BII) \end{bmatrix}$$

Logit scale:

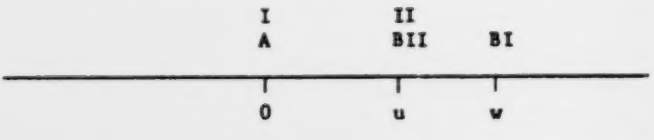


Figure 2.6 Saltus Matrix and Logit Scale for Case (vi)

d_{BI}, is in particular need of interpretation. If it were impossible for a person in group I to succeed on an item of type B, then d_{BI} should be infinity. However, at the empirical level, where a hierarchical theory meets reality, many factors can combine to make d_{BI} finite - guessing, copying, carelessness on item type A, and error in recording of results, are just a few. As d_{BI} nears d_{BII}, the question must be asked whether the asymmetry of the Saltus matrix is sufficient to claim that the rigidity is important. This question can be answered only with respect to the substantive application.

Saltus maintains the probabilistic nature of the Rasch model, and the additive interpretation of the parameters. The role of local independence is maintained, but the range of person-item behaviour that can be modelled has been expanded. Asymmetric patterns, such as those represented in Case (iv), that would constitute breaches of local independence under the Rasch model, are included in the Saltus model. These patterns can be detected within the Rasch model by the analysis of misfit statistics as described above.

Separation of person and item estimates is also maintained. By considering conditional probabilities, the person parameters can be eliminated, leaving only item and Saltus parameters. Similarly, item parameters can be eliminated, leaving the person and Saltus parameters, and Saltus parameters can be eliminated, leaving only the person and item parameters. The scale is still an equal interval scale using the logit as the natural unit. The minimal set of sufficient statistics has enlarged to incorporate the Saltus parameters.

One important difference between the Rasch model and the Saltus model is that they place their largest standard errors in different parts of the logit scale. The Rasch model places its largest standard errors at the extreme scores for the combined person groups and item types (i.e. at zero and the maximum score), and its smallest in the middle. Saltus places its largest standard errors at the extreme scores for each person group, so that quite large standard errors occur at the gap between the two person groups. Given that the assumption of rigidity is true, it seems reasonable to expect large errors over the gap. A person who has succeeded on all the items of type A, but failed all of type B, is teetering on the brink of the gap - one more success and that person would be classified in group II - and the magnitude of this potential change is expressed as a large standard error in the Saltus model.

Relationship to the Generic Theory

The two attributes of the generic theory of hierarchical development that must be understood in relation to Saltus are rigidity and gappiness. Gappiness is a property of the substantive content of the items: no 'in-between' or 'transition' stages are to be

represented amongst the items. This is indicated in both the Rasch model and in Saltus by segmentation of the logit scale. Saltus adds to this the ability to measure rigidity through the asymmetry index. In order to investigate the rigidity between two stages, they must be theoretically distinguished as having gappiness and this gappiness must have been demonstrated through the segmentation index. Segmentation is an expression of the separation of the content into separate stages: asymmetry is an expression of the directionality of development.

Segmentation is measured through two segmentation indices - one for each person group. The person group I and person group II segmentation indices are

$S_I = (\gamma_{Bmin} - \gamma_{BI}) - (\gamma_{Amax} - \gamma_{AI})$, and

$S_{II} = (\gamma_{Bmin} - \gamma_{BII}) - (\gamma_{Amax} - \gamma_{AII})$ respectively.

Note that when the gaps are equal, so that the asymmetry index is zero, the group segmentation indices are also equal. Segmentation has been demonstrated when these indices are greater than a chosen standard: for estimates this could be two or three times the standard error of the item calibrations. The strength of the segmentation is indicated by how much greater the segmentation indices are than this standard. The segmentation indices are not of equal importance. Rigidity focusses attention on the inability of persons in group I to succeed on items of type B, and so the group I segmentation index is more crucial. It is possible to imagine a theory of hierearchical development in which the difference in difficulty between type A and type B items is reduced to nothing as a person masters stage B. In this case, S_{II} would be very small, but the segmentation would still be expressed by S_I because that is the measure of the learning that must be done in order to get to the higher stage.

Asymmetry is expressed in the Saltus model through the asymmetry index D. Generally we expect that, while persons operating at stage B might fail items of type A at some rate determined by the many factors covered by the term 'human error', persons operating at stage A cannot succeed on items of type B except through some non-cognitive strategy such as guessing or cheating. Tests of cognitive development are (or should be) designed to minimise these non-cognitive strategies. Thus, one would expect that d_{BI} would be greater than d_{BII}, so the asymmetry index should be positive. If the asymmetry index is positive, the group I segmentation index must be greater than the group II segmentation index. If the asymmetry index is zero, then segmentation might still be present, indicating that item type B is harder than item type A, but there would be no indication of a distinct change in perspective associated with a stage transition. If the asymmetry index is negative, then the type B items would be harder for the group II people than for the group I people: one could speculate that this situation could arise if certain item types provoked guessing in the ignorant and misled the able. But this type of situation is difficult to reconcile with the concept of a

developmental hierarchy - if it were observed, then one would seek an explanation in terms of flawed items or faulty theory.

Estimation of Parameters

To estimate the parameters in the Saltus model, the unconditional maximum likelihood procedure, UCON, is used (Wright and Panchepakesan, 1969), with adaptations for the Saltus parameters; this adaptation is called the UCONG procedure. For person i with ability β_i, item j with difficulty δ_j, and Saltus parameter γ_{hk}, where h is a function of i and k is a function of j, the probability of a response y_{ij}, is

$$P(y_{ij} \mid \beta_i, \delta_j, \gamma_{hk}) = \frac{\exp(y_{ij}(\beta_i - \delta_j + \gamma_{hk}))}{1 + \exp(\beta_i - \delta_j + \gamma_{hk})} \quad .$$

Using local independence, the likelihood of the data matrix $((y_{ij}))$ is modelled as the continued product of the unconditional probabilities over L items and N persons:

$$\Lambda = \prod_{i=1}^{N} \prod_{j=1}^{L} \frac{\exp(y_{ij}(\beta_i - \delta_j + \gamma_{hk}))}{1 + \exp(\beta_i - \delta_j + \gamma_{hk})}$$

$$= \frac{\exp(\sum_i \sum_j y_{ij}(\beta_i - \delta_j + \gamma_{hk}))}{\prod_i \prod_j (1 + \exp(\beta_i - \delta_j + \gamma_{hk}))} \quad .$$

The log-likelihood is:

$$\lambda = \sum_{i=1}^{N} r_i \beta_i - \sum_{j=1}^{L} s_j \delta_j + \sum_{k=A}^{B} \sum_{h=I}^{II} t_{hk} \gamma_{hk}$$

$$- \sum_{i=1}^{N} \sum_{j=1}^{L} \log(1+\exp(\beta_i - \delta_j + \gamma_{hk})) \qquad (8)$$

where $r_i = \sum_j y_{ij}$ is the score for each person,

$s_j = \sum_i y_{ij}$ is the score for each item, and

$t_{hk} = \sum_{G(h,k)} y_{ij}$ is the Saltus score,

that is, the score for all items of type k over all persons in group h, and G(h,k) is the set of all person-item pairs in which the person group is h and the item type is k.

Equation (8) is the logarithmic version of the required form for conditionally sufficient statistics (Kendall and Stuart, 1969, p.9), for each of the parameters. Thus, the item scores, given the person abilities and the Saltus parameters, are sufficient for the item difficulties; the person scores, given the item difficulties and the Saltus

parameters, are sufficient for the person abilities; and the Saltus scores, given the person abilities and the item difficulties, are sufficient for the Saltus parameters.

The estimates which maximize the likelihood are the same as those which maximize the log-likelihood. The maximum of the log-likelihood is given by the set of parameters for which the partial first derivatives are zero and the partial second derivitives are negative. The maximum likelihood estimates are the solutions to the equations:

$$\begin{aligned} &r - \textstyle\sum_j \Pi_{ij} = 0, \text{ for } r = 1, \ldots, L-1, \\ &-s_j + \textstyle\sum_i \Pi_{ij} = 0, \text{ for } j = 1, \ldots, L \\ &\text{and } t_{hk} - \textstyle\sum_{G(h,k)} \Pi_{ij} = 0, \text{ for } h=I,II \text{ and } k=A,B \ . \end{aligned} \tag{9}$$

These equations are solved under the constraint that, for each Saltus parameter, the average of the person score estimates and the average of the item difficulty estimates are set to zero. This generalizes the situation for the Rasch Model where the item and person parameters were held constant within types and groups. It ensures that the relationship between the groups of persons and the item types is measured only by the Saltus parameters. If b_r is the estimate of β_r, d_j is the estimate of δ_j, and we adopt the convention that I is the set of scores for person group I, II is the set of scores for person group II, A is the set of subscripts for items of type A and B is the set of subscripts for items of type B, then the constraints can be written

$$\begin{aligned} &\textstyle\sum_I b_r = 0 \\ &\textstyle\sum_{II} b_r = 0 \\ &\textstyle\sum_A d_j = 0 \text{ and} \\ &\textstyle\sum_B d_j = 0. \end{aligned} \tag{10}$$

Equations (9), under the constraints (10), are solved using Newton's technique. If $b_r^{(t)}$, $d_j^{(t)}$ and $g_{hk}^{(t)}$ are estimates of the person, item and Saltus parameters, respectively, and $P_{rj}^{(t)}$ is the probability of a person with score r getting item j correct calculated using these estimates, then the estimates will be improved by:

$$b_r^{(t+1)} = b_r^{(t)} - \frac{r - \sum_{j=1}^{L} P_{rj}^{(t)}}{-\sum_{j=1}^{L} P_{rj}^{(t)}(1 - P_{rj}^{(t)})}$$

$$\text{for } r = 1, \ldots, L-1,$$

$$d_j^{(t+1)} = d_j^{(t)} - \frac{-s_j + \sum_{r=1}^{L-1} N_r P_{rj}^{(t)}}{-\sum_{r=1}^{L-1} N_r P_{rj}^{(t)}(1 - P_{rj}^{(t)})}$$

for $j = 1, \ldots, L$

where N_r is the frequency of score r

$$g_{hk}^{(t+1)} = g_{hk}^{(t)} - \frac{t_{hk} - \sum_{j \in k} \sum_{r \in h} N_r P_{rj}^{(t)}}{-\sum_{j \in k} \sum_{r \in h} N_r P_{rj}^{(t)} (1-P_{rj}^{(t)})}$$

for h=I,II and k=A,B .

Asymptotic standard errors can be estimated from the denominator of the last iteration:

$$SE(b_r) = (\textstyle\sum_j P_{rj}(1-P_{rj}))^{-\frac{1}{2}}$$

$$SE(d_j) = (\textstyle\sum_r N_r P_{rj}(1 - P_{rj}))^{-\frac{1}{2}}$$

$$SE(g_{hk}) = (\sum_{j \in k} \sum_{r \in h} N_r P_{rj}(1 - P_{rj}))^{-\frac{1}{2}},$$

where P_{rj} is the probability found using the final estimates.

Note that when the sufficient statistics s_j or t_{hk} are zero or maximal, a solution is not attainable. If an item score is zero or maximal, the analysis can be carried out without that item. If a Saltus score is zero or maximal, the asymmetry index is not obtainable from the data. Data sets where this occurs are called 'intractable'.

The UCON estimation procedure has been found to give results which are slight overestimates of the parameters. This is due to the use made of the person estimates at each iteration of the item estimates, as though they were parameters. The correction made for this bias is to deflate the item difficulty estimates by (L-1)/L, and then re-estimate the person abilities (Wright, Masters and Ludlow, 1981). For Saltus, a similar correction is made, based on the number of items within each stage. Items of type A are corrected by $(L_A-1)/L_A$, and items of type B are corrected by $(L_B-1)/L_B$. For gaps, the correction is $(L^*-1)/L^*$, where L^* is the average of L_A and L_B.

An initial approximation is made using a modification of the PROX technique (Wilson, 1984 pp.81-84).

Assessing the Fit of Data to the Model

The fit of data to the Saltus model is examined through the use of two statistics. The logit bias (Wright, 1982) for an item with respect to a person group is the average amount by which the estimates underestimate the success of that group on the item, in logit units. The standardized bias measures the same thing scaled to have a mean of 0 and a standard deviation of 1.

For N_I persons of type I, with y_{ij} representing the observed score of person i on item j and P_{ij} the expected score of person i on item j, the logit bias of item j with respect to group I is

$$H_I(j) = \frac{\sum_{i=1}^{N_I} (y_{ij} - P_{ij})}{\sum_{i=1}^{N_I} v_{ij}}$$

where $v_{ij} = P_{ij}(1 - P_{ij})$ is the item variance.

The logit bias gives a measure, in logit units, of how much on average the model has underestimated the difficulty of an item with respect to a particular person group. For example, a group I logit bias of -1.00 on an item with estimated difficulty 1.00 logits, indicates that, for an average group I person, the item's difficulty was overestimated by 1.00 logits, or, alternatively, that group I ability was underestimated, for that particular item, by 1.00 logits.

The standardized bias of item j with respect to group I is

$$G_I(j) = \sum_{i=1}^{N_I} z_{ij}/(N_I)^{\frac{1}{2}}$$

where $z_{ij} = (y_{ij} - P_{ij})/(v_{ij})^{\frac{1}{2}}$ is the standardized residual (Wright, 1982).

Because it incorporates a measure of the underlying standard error of measurement, the standardized bias gives perspective to the corresponding logit bias: it has expected value 0 and standard deviation approximately 1. A standardized bias of less than 1 indicates that, no matter how large is the logit bias, it is small compared to the variation expected, and so, should not be interpreted. A standardized bias of 2 or more indicates that the corresponding logit bias should be investigated.

These formulae can be repeated using group II to give the logit and standardized biases for group II, represented by $H_{II}(j)$ and $G_{II}(j)$.

For persons, the same procedure can be repeated with respect to item types. For item type k, with L_k items, the logit bias is

$$H_k(i) = \frac{\sum_{j=1}^{L_k} (y_{ij} - P_{ij})}{\sum_{j=1}^{L_k} v_{ij}} \qquad \text{for } k=A,B,$$

and the standardized bias is

$$G_k(i) = \sum_{j=1}^{L_k} z_{ij}/(L_k)^{\frac{1}{2}} \qquad \text{for } k=A,B.$$

The interpretation of these quality control statistics is analogous to the interpretations for the items.

Checking the Performance of Saltus

A series of simulations was conducted to detect bias in the model (Wilson, 1984, pp.86-125). These showed that when the asymmetry index was zero, the generators were accurately estimated. As the asymmetry index increased the estimates remained good except when the group II gap was small (less than 2) and the group I gap was large (more than 4), in which case the group I gap was under-estimated: this was found to be associated with 'cross-overs' - persons who had been classified by their scores into the wrong person group. Simulations are expensive and time-consuming to conduct: Wilson (1984, pp.152-9) describes an approach based on tailored simulations, but a better solution is needed.

CHAPTER 3

A SUBTRACTION HIERARCHY

The Subtraction Tasks

A large set of constructed-response subtraction items was developed according to a Gagnean learning hierarchy at the Australian Council for Educational Research: they were later published as the 'RAPT in Subtraction' tests (Izard et al., 1983). The items were tried with a structured probability sample of students in Years 3 and 4 at schools in Victoria and New South Wales. The students were selected as intact classrooms sampled from the population of schools structured by: geographical location (rural, suburban, inner urban), size (large, small), and type of school administration (state, Catholic, private). After several rounds of revision of item objectives, and after some further items were written and tested, the researchers settled on the hierarchy shown in Table 3.1 as the best respresentation of the subtraction learning sequence.

The items were analysed using the Rasch model and the results are summarized in Figure 3.1; the RAPT units used are a linear transformation of the original logits. Each objective is represented by six items; although a sequential development through the objectives is clearly indicated, segmentation is not evident except between objectives 3 and 4.

Compare the definitions of objectives 1 and 5 with those for 2 and 6. These two pairs of objectives both test the regrouping step, the first with 2-digit items, the second with 3-digit items. These two pairs provide an interesting way to replicate the step from not being able to regroup to the attainment of the regrouping skill. The analysis will focus on this part of the hierarchy because it clearly demonstrates segmentation and because of the added interpretation made possible by the repetition of the regrouping step across different numbers of digits. The subsample used in the analyses is described in Table 3.2. All items and students used in the analysis are from the original sample because the additional students and items used in the final construction of the RAPT tests were not available. The items are given in Figure 3.2. In the following analyses, the items without regrouping (i.e. those from tests 1 and 2) constitute the type A items, and those with regrouping (i.e. those from tests 5 and 6) constitute the type B items. The Rasch analysis of the subtraction items has demonstrated that the items segment the logit scale. The Saltus analysis will allow the exploration of this segmented scale for the asymmetry that indicates a rigid hierarchy.

In the table and figure captions, the analysis concerned will be indicated by symbols for the number of digits in the subtraction items (2 or 3), the state (V or NSW), the Year of the students (3 or 4) and the sex of the students (M or F). Thus, '3V4' is an

Table 3.1 Subtraction Objectives

Test number	Objective
1	Subtract a 2-digit subtrahend from a 2-digit minuend, with no regrouping.
2	Subtract a 3-digit subtrahend from a 3-digit minuend, with no regrouping.
3	Subtract a 1-digit subtrahend from a 2-digit minuend, with no regrouping.
4	Subtract a 1-digit subtrahend from a 2-digit minuend, with regrouping.
5	Subtract a 2-digit subtrahend from a 2-digit minuend, with regrouping.
6	Subtract a 3-digit subtrahend from a 3-digit minuend, with regrouping from one place and no zeroes in the minuend.
7	Subtract a 3-digit subtrahend from a 3-, 4, or 5-digit minuend, with regrouping from two places and no zeroes in the minuend.
8	Subtract a 3- or 4-digit subtrahend from a 3-, 4- or 5-digit minuend, with regrouping where necessary and zeroes in the minuend.

analysis of the 3-digit item data from the Victorian Year 4 students, and '2NSWF' is an analysis of the 2-digit item data from the female New South Wales students.

Comparison with the Rasch Analysis

The sample of 75 Year 4 students from Victoria taking the 3-digit items was chosen as the main comparison group because the grade 3 groups were found to be less stable in their performances and the New South Wales fourth year students gave an intractable set of results on the 2-digit items.

The Rasch item estimates are given in Table 3.3 and are illustrated in Figure 3.3 in this table and figure the scale has been centred on the average of the A items to match the procedure for the Saltus analyses. The items show segmentation: the segmentation index (the difference between the hardest type A item and the easiest type B item) is 1.19, considerably greater than the 0.22 root mean square of the standard errors of the inner two items. The fit statistics show the characteristically large negative values above the suspected gap.

The Saltus item estimates are given in Table 3.4 and are illustrated in Figure 3.4. The estimated Saltus matrix is

$$\begin{bmatrix} 0.44\ (AI) & 1.88\ (AII) \\ -4.55\ (BI) & 0.55\ (BII) \end{bmatrix}$$

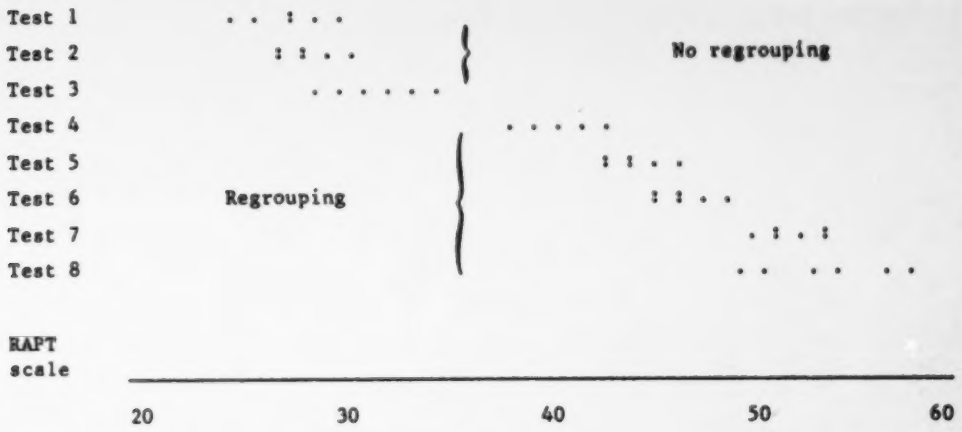


Figure 3.1 Item Difficulties on the RAPT Subtraction Scale

with standard errors

$$\begin{bmatrix} 0.35 & 0.23 \\ 0.72 & 0.13 \end{bmatrix}.$$

The asymmetry index is 3.61, with standard error 0.84: the negative fit statistics in the Rasch analysis have accurately indicated a large asymmetry. Note that the distance in Figure 3.3 of 2.36 between the average of the group A and group B items for the Rasch analysis has been decomposed into a distance of 1.38 for group II and 4.99 for group I. The width of the item sets has not altered much: 1.47 and 0.82 for the type A and B items under the Rasch analysis, and 1.18 and 0.90 under the Saltus analysis. The positive asymmetry index is illustrated in Figure 3.4 by the distance between BII and BI. The Rasch analysis has 'averaged' the two locations of B given by Saltus. That is, Saltus has shown that, for the group I students, the type B items are further from the type A items than was indicated by the Rasch analysis, and that, for the group II students, the type B items are closer to the type A items than was indicated by the Rasch analysis.

The Rasch and Saltus estimates for the type A items are plotted against one another in Figure 3.5. This figure shows that the two models are calibrating the A items in the same way. The same comparison for item type B is shown in Figure 3.6: there are two locations for each item in this figure, one for group I and one for group II. Within each group, the items fall on a line parallel to the identity line, indicating that the two models are giving the same relative difficulties between the items. However, for group I, Saltus has placed the B items 1.00 logits below the Rasch location, and, for group II, Saltus has placed the B items 2.61 items above the Rasch location. The asymmetry index measures the distance between the location of the B items for the two groups (i.e. 3.61 = 1.00 + 2.61).

Table 3.2 Sample Used for Saltus Analyses

State	2-digit items		3-digit items	
	Year 3	Year 4	Year 3	Year 4
Victoria	42 boys	27 boys	25 boys	36 boys
	41 girls	24 girls	23 girls	39 girls
Total (Vic.)	83	51	48	75
New South Wales	25 boys	14 boys	37 boys	17 boys
	38 girls	17 girls	30 girls	22 girls
	18 unknown	21 unknown	29 unknown	26 unknown
Total (NSW)	81	52	96	65
Grand Total	164	103	144	140

The fit statistics for the Saltus model (given in Table 3.4) are not as large as those for the Rasch model (given in Table 3.3), nor do they show a discernible pattern. This lack of pattern and small size for the fit statistics implies that a better fit has been obtained.

The standard errors for the item difficulties on the logit scale are derived from the item standard errors from the UCONG estimation procedure, s_i, and the standard errors of the gaps, s_{kh}, which are used to locate the items on the scale. For group I students, the location of item i is d_i, if i is of type A and d_i+d_{BI}, if i is of type B, where d_i is the estimated item difficulty for item i and d_{BI} is the estimated group I gap. Hence, the standard errors for student group I (I shall use 'student' rather than 'person' throughout this chapter) are given by

$s'_i{}^2 = s_i{}^2$ if i is of type A, and

$s'_i{}^2 = s_i{}^2 + s_{BI}{}^2$ if i is of type B

where s_i is the estimated standard error for item i,
and s_{BI} is the standard error of the group I gap.

Similarly, for students in group II, the standard error of item i is given by

$s'_i{}^2 = s_i{}^2$ if i is of type A, and

$s'_i{}^2 = s_i{}^2 + s_{BII}{}^2$ if i is of type B

where s_{BII} is the standard error of the group II gap.

The standard errors for group I students attempting type B items are largest because they are far from those items. The standard errors for the group II students attempting type B items are smaller because they are close to those items. The errors for the A items are approximately the same as those for the Rasch analysis. The errors for the B

Test 1: 2-digits, no regrouping

1.1	99 -48	1.2	73 -42	1.3	37 -14
1.4	64 -22	1.5	98 -74	1.6	75 -54

Test 2: 3-digits, no regrouping

2.1	598 -123	2.2	678 -234	2.3	997 -411
2.4	364 -221	2.5	369 -145	2.6	689 -352

Test 5: 2-digits, regrouping

5.1	71 -48	5.2	73 -68	5.3	81 -65
5.4	23 -15	5.5	32 -18	5.6	55 -17

Test 6: 3-digits, regrouping

6.1	417 -126	6.2	455 -173	6.3	826 -452
6.4	352 -119	6.5	565 -384	6.6	865 -639

Figure 3.2 Items in the Subtraction Tests

Table 3.3 Rasch Estimates for the 3V4 Sample

Item type	Item label	Estimated Difficulty	Estimated Error	Fit
B	6.1	2.87	0.31	-0.41
	6.6	2.53	0.30	-1.13
	6.3	2.38	0.29	-1.52
	6.4	2.30	0.29	-1.80
	6.2	2.14	0.28	-1.16
	6.5	2.05	0.17	-2.06
A	2.6	0.86	0.39	0.57
	2.5	0.34	0.45	1.12
	2.1	-0.03	0.51	-0.10
	2.3	-0.29	0.55	0.61
	2.2	-0.29	0.55	0.92
	2.4	-0.61	0.62	0.36

items are larger than those for the Rasch analysis because the scale has stretched to accommodate the more accurate estimation of the Saltus model. The coefficient of variation gives a measure of the impact of this change in standard errors: the average coefficient of variation for the Rasch analysis for the B items is 4.15, and for the Saltus analysis it is 3.14 for group I and 5.74 for group II. Thus the realtive accuracy of the Rasch estimates is between the two for the Saltus estimates, with the group II estimate performing best, as one might expect since group II has been constructed to match item type B.

The Rasch and Saltus score estimates are given in Table 3.5 and illustrated in Figures 3.3 and 3.4. The count of scores shows a bi-modal pattern that is typical where there is rigidity between item types. The Rasch estimates for the scores progress monotonically as one expects for items measuring a single attribute - the higher the score, the higher the logit ability. For Saltus, however, this is not so. The score

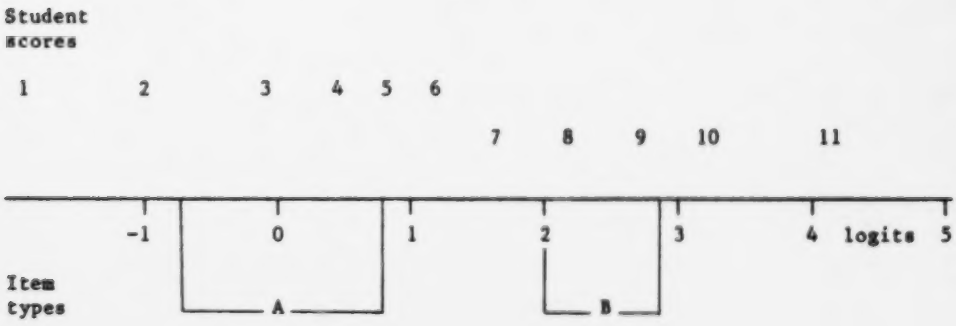


Figure 3.3 Rasch Estimates for the 3V4 Sample

Table 3.4 Saltus Estimates for the 3V4 Sample

Item type	Difficulty order	Group I		Group II		Fit
		Difficulty	Error	Difficulty	Error	
B	6.1	5.52	0.85	1.91	0.38	0.49
	6.6	5.16	0.86	1.55	0.40	-0.19
	6.3	5.00	0.86	1.39	0.41	-0.43
	6.4	4.91	0.87	1.30	0.42	-0.58
	6.2	4.73	0.87	1.12	0.43	0.27
	6.5	4.62	0.87	1.01	0.45	-0.62
A	2.6	0.68	0.36			-0.52
	2.5	0.31	0.42	Same		0.33
	2.1	-0.02	0.48	as		-0.28
	2.3	-0.23	0.53	group		0.25
	2.2	-0.23	0.53	I		0.34
	2.4	-0.50	0.60			0.18

estimates fold back onto themselves, so that a group II student with score 10 gets a slightly lower estimate than a group I student with score 6!

The reason for this is that the logit scale has been set up with the A items as the reference point. That a group I student with score 6 has an estimated position 0.28 logits above a group II student with score 10, means that a group I student has a slightly higher probability of getting the A items correct than the group II student who scored 10. However, as the student who scored 6 is in group I, that student's chance of getting a type B item correct is very small, as indicated by the position of BI at 4.99 logits in Figure 3.4, whereas the group II student who scored 10 has quite a good chance of getting a type B item correct, as indicated by the position of BII at 1.38 logits. These probabilities are detailed in Table 3.6. This table shows the large difference in the difficulty of type B items for student groups I and II: it also shows that the type B items,

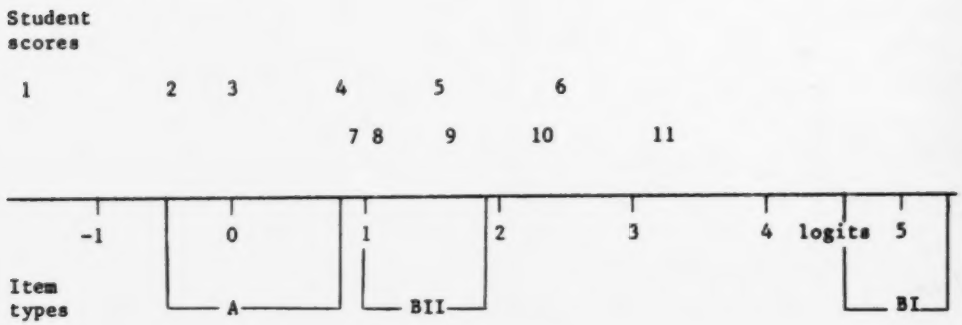


Figure 3.4 Saltus Estimates for the 3V4 Sample

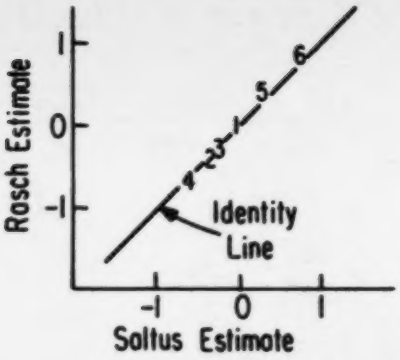


Figure 3.5 Rasch and Saltus Difficulties for Type A Items

even though their logit range of 1.18 is the same for both student groups, have a much smaller range of probability for the group I students than for the group II students.

Consider, for example, a group I student who scored 4 and a group II student who scored 7. These have approximately the same logit abilities, that is, the type A items look much the same to both: the group I student has a probability of success ranging from 0.51 to 0.77 and the group II student has a probability of success ranging from 0.55 to 0.88. This can be interpreted as meaning that a student who has almost mastered the non-regrouping items, but has had little or no success on the regrouping items, has much the same ability with respect to the non-regrouping items as a student who has just begun to succeed on the regrouping items. This implication fits well with the type of item involved - correct solution of these subtraction problems requires sustained

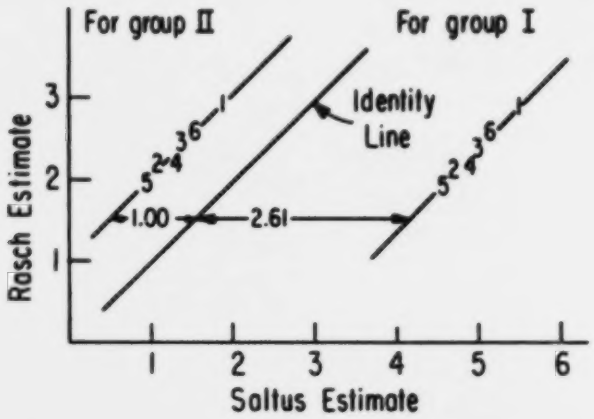


Figure 3.6 Rasch and Saltus Difficulties for Type B Items

Table 3.5 Score Estimates for the Rasch and Saltus Analyses

Student group	Score	Count	Rasch		Saltus	
			Ability	Error	Ability	Error
II	11	38	2.95	1.08	3.14	1.08
	10	16	2.08	0.83	2.31	0.83
	9	5	1.48	0.74	1.75	0.74
	8	2	0.97	0.78	1.30	0.69
	7	2	0.49	0.69	0.90	0.66
I	6	6	-0.02	0.69	2.59	1.09
	5	3	-0.45	0.69	1.53	1.02
	4	3	-0.95	0.71	0.73	0.92
	3	0	-1.48	0.75	0.04	0.89
	2	0	-2.10	0.84	-0.66	0.94
	1	0	-2.99	1.09	-1.59	1.15

concentration on an algorithm combining several entry-level skills such as remembering subtraction tables and keeping columns aligned. These skills are still under improvement while the student is being introduced to new topics such as regrouping. Contrast this with the Rasch results, which give a student who scored 7 a much higher chance of success on every item than a student who scored 4, and hence a much higher chance of success on the non-regrouping items. Because the two types of items have not been identified in the Rasch model, the detail of the different student behaviours for the different item groups is completely missing.

For the regrouping items, however, the difference in ability is very noticeable. The group II student is finding the items difficult, he is having reasonable success (0.50) at the easier ones, and just a little success at the harder ones (0.24). But for the group I student these items are, uniformly, almost impossible - the probability of success ranges down from 0.12 to 0.00. This reduction in the range of probabilities illustrates the

Table 3.6 Probability of Success on Easiest and Hardest Items

Group	Score	Probability of Success			
		Non-regrouping items		Regrouping items	
		Easiest	Hardest	Easiest	Hardest
I	1	0.25	0.09	0.00	0.00
	2	0.46	0.21	0.01	0.00
	3	0.61	0.32	0.01	0.00
	4	0.77	0.51	0.02	0.01
	5	0.89	0.71	0.05	0.02
	6	0.96	0.87	0.12	0.05
II	7	0.80	0.55	0.50	0.24
	8	0.86	0.65	0.60	0.32
	9	0.90	0.74	0.70	0.42
	10	0.94	0.84	0.81	0.56
	11	0.97	0.92	0.91	0.75

Table 3.7 Saltus Item Fit Statistics for the 3V4 Sample

Item	Student Group I		Student Group II	
	Logit bias	Standardized bias	Logit bias	Standardized bias
2.1	-0.89	-0.64	0.44	0.85
2.2	1.21	1.23	-0.56	-0.84
2.3	0.37	0.56	-0.15	-1.01
2.4	0.10	0.32	-0.03	-0.66
2.5	0.22	0.01	-0.06	-0.30
2.6	-0.36	-0.22	0.17	0.37
6.1	6.11	1.76	-0.04	-0.14
6.2	1.86	2.72	-0.05	-0.08
6.3	-1.03	-0.49	0.05	0.23
6.4	-1.03	-0.52	0.06	0.34
6.5	-1.05	-0.61	0.08	0.31
6.6	-1.03	-0.44	0.04	0.19

difference in perspectives between the two groups - the regrouping items are near enough to the group II students for the details of their construction to make an observable difference in the students' performance, but the overwhelming difficulty of mastering regrouping has pushed the items so far beyond the ability of the group I students that the differences between the items have become insignificant.

Another way in which the Rasch analysis differs from the Saltus analysis is in the pattern of the standard errors of the score parameters given in Table 3.5. For the Rasch analysis they increase towards the more extreme scores, but for the Saltus analysis they increase towards the extreme scores in each score group. The important difference is in the middle: for Rasch, this is where the smallest standard errors are, but for Saltus, large standard errors occur here because this is the critical region between the two student groups. A change of just one score here could put a student into the other group, resulting in a great deal of change in the modelled probabilities of success for that student.

The item fit statistics given in Table 3.7 for the Saltus analysis draw attention to items 6.1 and 6.2, which have standardized biases for student group I of 1.76 and 2.72, and logit biases of 6.11 and 1.86. The logit biases indicate that the group I students exhibited more success (6.11 and 1.86 logits worth) on the items than the estimated parameters would indicate, and the standardized biases indicate that these logit biases are large compared to the level of variability that is expected in a probabilistic model. Since the performance of the group II students would dominate the estimation of the type B items, some group I students are doing relatively better on items 6.1 and 6.2 (compared to 6.3 to 6.6) than the group II performances would predict. As these items are at the beginning of the test form, perhaps some students who did not have the regrouping skill had time to apply a 'low-stress' algorithm (such as counting) to these first few items, thus circumventing the intent of the test-makers.

Table 3.8 Saltus Student Fit Statistics from the 3V4 Sample

Student type	Student No.	Item type A Logit bias	Item type A Std. bias	Item type B Logit bias	Item type B Std. bias	Responses to items: Test 2						Test 6					
						1	2	3	4	5	6	1	2	3	4	5	6
'Careless Error' (8 students)	60	-4.37	-2.81	1.18	0.98	1	1	1	0	1	1	1	1	1	1	1	1
'Misclassification?' (1 student)	58	-1.48	-0.76	3.67	2.52	1	1	1	1	0	1	1	0	0	0	0	0
'Inconsistent with ordering of item types' (1 student)	69	-2.47	-2.04	1.34	1.62	1	0	1	1	0	0	1	1	1	1	0	1
'Fit Saltus' (65 students)	14	1.09	0.66	-1.04	-0.47	1	1	1	1	1	1	0	0	0	0	0	0

The student fit statistics for the Saltus analysis, some of which are given in Table 3.8. draw attention to two sets of students who are not fitting the estimated model well. The first set is composed of students who scored 11, placing them in group II, but who were unsuccessful on one non-regrouping item. A typical case is student No. 60 who, with an estimated ability of 1.26, failed item 2.4, causing a standardized bias of -2.81 for the non-regrouping items. There were eight students in this set, and the result may be considered to indicate careless error on the part of these students. The second set consists of just two students. Student No.58 scores 6, but he did not conform to expected behaviour because he was successful on item 6.1 but unsuccessful on item 2.5, causing a standardized bias of 2.52 on the regrouping items. The categorization employed by Saltus has placed him in group I, so that the success on item 6.1 seems surprising, but he could equally be considered a low-ability group II student, which would make the failure on item 2.5 surprising. This case would bear further investigation if the student were available for interviewing. Student No.69 failed on items 2.2, 2.5, 2.6 and 6.5, and succeeded on the rest. She was the only student in the sample who succeeded on more regrouping items than non-regrouping items. She was also the only student who caused large misfit statistics for both the Rasch and Saltus analyses; a total fit of 3.00 for the Rasch model, and standardized biases of -2.04 and 1.62 for item types A and B for the Saltus model. The inconsistency of this student is difficult to interpret and would also have to be investigated through interview.

One final comparison can be made between the two analyses. It has been left to last because of the emphasis that has been placed on the interpretation of results rather than on their statistical features. This is the statistical improvement resulting from the Saltus model. Because the Saltus model is the same as the Rasch model with the addition of one asymmetry parameter, and because the same sample was used for both analyses, a likelihood ratio test can be performed to compare the fit of the two models. The total log-likelihood for the Rasch analysis was -307.91 and for the Saltus analysis, -288.95. This gives a likelihood ratio chi-squared statistic of

$$-2(-307.91 + 288.95) = 37.92$$

on 1 degree of freedom, which is significant at the 0.001 level. This indicates that the extra degree of freedom used by the Saltus model for its Saltus parameter makes a significant improvement in fitting the data.

Replicating the Saltus Results

The complexity of the subtraction sample allowed some attempt at quasi-experimental design. The geographical structure permitted a comparison between the Victorian and New South Wales results, the grade structure permitted a comparison between Year 3

Table 3.9 Subtraction Saltus Matrices, with Standard Errors in Parentheses

	Victoria				New South Wales			
	2-digit		3-digit		2-digit		3-digit	
Year 3	0.61	2.86	0.67	1.97	0.43	2.60	0.33	2.87
	(0.21)	(0.39)	(0.40)	(0.29)	(0.22)	(0.32)	(0.19)	(0.26)
	-7.01	0.22	-7.16	0.48	-4.51	0.25	-3.93	0.20
	(1.00)	(0.16)	(1.00)	(0.18)	(0.34)	(0.15)	(0.37)	(0.12)
Year 4	0.31	1.92	0.44	1.88			0.31	2.09
	(0.39)	(0.29)	(0.35)	(0.23)			(0.45)	(0.27)
	-3.88	0.49	-4.55	0.50			-3.50	0.32
	(0.60)	(0.16)	(0.72)	(0.13)			(0.52)	(0.13)

and 4 results, and the structure of the item objectives permitted a comparison between different realizations of the same step in the hierarchy. In addition, an investigation of male-female differences was made, but due to insufficient numbers and some missing data in the New South Wales sample, this could only be studied by pooling the year levels of the Victorian sample. Each of these comparisons constitutes one small step along the way to fully understanding the meaning of a gap: the specification of those item features which do not alter a gap, those which change its measure but not its quality, and those which change it into something other than a gap, would delineate a realization of a stage transition within a generic theory of hierarchical development. The discussion of these results is confined to the Saltus parameters, as the details for the other parameters were similar to those for the Saltus analysis just discussed.

The Saltus parameters for the 2- and 3-digit items over the Year 3 and 4 Victorian and New South Wales samples are given in Table 3.9 and the asymmetry indices are given in Table 3.10. The Saltus parameters are presented as Saltus matrices. The 2-digit results for the New South Wales Year 4 students are not given as this sample proved intractable to Saltus analysis. Consider first the Victorian Year 4 sample (the 3-digit case was discussed in detail in the previous section). These two matrices show a similar pattern: all except the two BI Saltus parameters (-3.88 for 2-digit items and -4.55 for 3-digit items) are close, but even they are less than one standard error apart. Thus, for

Table 3.10 Subtraction Asymmetry Indices, with Standard Errors in Parentheses

	Victoria		New South Wales	
Year	2-digit	3-digit	2-digit	3-digit
3	4.99	6.34	2.59	1.59
	(1.11)	(1.13)	(0.54)	(0.53)
4	2.76	3.61		2.04
	(0.79)	(0.84)		(0.75)

the Victorian Year 4 sample, the pattern of results, discussed above for the 3-digit case, is repeated for the 2-digit case. This indicates that, for this sample at least, the regrouping gap is not much affected by the different number of digits used in these items.

Consider the Victorian Year 3 sample. Once again the pattern of results of the 2- and 3-digit items are similar. The AII Saltus parameters (2.86 for 2-digit items and 1.97 for 3-digit items), however, are somewhat more than one standard error apart. This seems reasonable, since one might expect less experienced Year 3 students to find the superficial differences between the 2- and 3-digit items more of a problem than the Year 4 students. This difference is interesting - the Year 3 students who are in group I are relatively better (about 1 logit) at the 2-digit type A items than the 3-digit type A items. This is consistent with the usual order of introduction of such problems. Moreover, these Year 3 students in group II are also about 1 logit better than the Year 4 students in group II at the type A items. (Compare the AII Saltus estimate for the Victorian Year 3 2-digit case, 2.86, with the AII Saltus estimate for the Victorian Year 4 2-digit case, 1.92.) This could be a case where a freshly honed skill deteriorates over the next year or so.

With some understanding of the differences between the 2- and 3-digit cases for the Victorian Year 3 students, attention can now be concentrated on the differences between Year 3 and Year 4. It is a noticeable difference, and is consistent across the two item types: the AI Saltus estimates are more negative for the Year 3 students (-7.01 and -7.16) than for the Year 4 students (-3.88 and -4.55). This means that the Year 3 group I students find the type B items even more difficult than the Year 4 students, which, given the relative lack of experience of the Year 3 students, is what we expect. Before making too much of this, however, notice that the logit parameters for the 3-digit case translate to a probability of success for a typical group I student attempting a typical group B item of 4.0×10^{-4} for the Year 3 students, and 6.8×10^{-3} for the Year 4 students. Although these probabilities are proportionally different, they are both so small that in the normal classroom setting they would look like 'never'.

Consider the New South Wales results. The 3-digit analysis on the New South Wales Year 4 students gave a pattern of results within a standard error of the corresponding Victorian sample. Unfortunately, the 2-digit analysis for the New South Wales Year 4 students was intractable. The Year 3 analyses are similar to the Year 4 analysis: there is no large increase in the BI Saltus estimate as in the Victorian sample. This could be due to an earlier introduction of the regrouping algorithm for brighter students in New South Wales.

The comparison of male with female samples was made difficult by the small number of cases within each of the cells when one sex was deleted, and also by the amount of missing data concerning the sex of the students in the New South Wales sample. So the New South Wales sample was left out and the year levels were collapsed

Table 3.11 Summary of Saltus Analyses for Victorian Boys and Girls

Sex	2-digit items			3-digit items		
	Saltus matrix		Asymmetry index	Saltus matrix		Asymmetry index
Boys	0.46 (0.23)	2.42 (0.35)	3.40 (0.73)	0.55 (0.35)	2.09 (0.25)	4.50 (0.84)
	-5.11 (0.58)	0.31 (0.16)		-5.61 (0.71)	0.43 (0.14)	
Girls	0.65 (0.31)	3.03 (0.60)	4.31 (1.23)	0.64 (0.42)	1.27 (0.57)	6.52 (1.31)
	-6.47 (1.00)	0.22 (0.22)		-6.28 (1.01)	0.87 (0.45)	

in the Victorian sample. The results for these collapsed samples are given in Table 3.11. Taking first the 3-digit items, the Saltus parameters indicate that the group I girls find type B items relatively harder than the boys, and the group II girls find the A items relatively harder than the group I boys. This is reflected in the difference in the asymmetry indices: 4.50 for boys and 6.52 for girls, which are between 1 and 2 standard errors apart. These results indicate that the girls are finding the regrouping gap more rigid than the boys. Compare this with the 2-digit item analyses. Here the group I girls once again find the type B items relatively harder than do the group I boys. But the group II girls find the type A items relatively easier than the group II boys, which makes the two asymmetry indices closer: 3.40 for boys and 4.31 for girls - the difference is not large compared to the standard errors of the asymmetry indices.

Overall, these results show a remarkable consistency. The patterns described in the previous section hold consistently across the replications described here; the exceptions have been found to have reasonable explanations. These explanations should not be seen as a way of 'making excuses' for discrepancies in the results, but rather as building blocks to a deeper and more detailed understanding of the conditions under which gaps occur in development and the stimuli that can reveal them.

CHAPTER 4

A RULE-ASSESSMENT HIERARCHY

Introduction

A rule-assessment approach to the understanding of cognitive development has been advanced by Robert S. Siegler (1981) as an adaptation of Piaget's theory. It attempts to assimilate two criticisms of the Piagetian position on developmental sequences. The first criticism concerns the sequence of development within a concept: Piagetians would evaluate this is in clinical interviews. Apart from the problems of the reliability of such procedures (Keating, 1980; Neimark, 1975), this practice has been found difficult to apply to young children (Bryant, 1974) and has been criticised on the grounds that children may possess a concept operationally but may not be able to articulate it (Brainerd, 1977). The second criticism concerns the sequence of development between concepts: Piagetian theory predicts certain synchronies in the development of the different concepts due to the over-arching effects of the stages. Empirical studies have shown that far more variation is present than the Piagetian literature would lead one to expect (Brainerd, 1978; Flavell, 1971; Keating, 1980; Neimark, 1975).

The most important characteristic of the rule-assessment approach is the specification of a series of increasingly powerful rules for solving problems. The behaviour of the learner is assumed to be dominated by the rule which he or she is using at a particular stage of development, and the sequence of development through the rules is assumed to be fixed. Thus far, it is basically the same as the Piagetian approach. It differs, however, in that it does not assume that these rules are the same across concepts, although the search for congruence between concepts consistutes a large and interesting part of the research. It also differs in that the data are collected as non-verbal choices to concrete problem-solving situations.

Siegler investigated the rule-assessment approach with three experimental situations involving proportionality: a balance scale task, a projection of shadows task and a probability task. For each, using task analysis, and by reference to previous empirical and theoretical work, a series of rules that children might use in tackling the task was hypothesized. Then a set of concrete problem types were developed which were easily replicable, which had a well defined set of variations and for which there were a small number of possible solutions so that the subject could indicate his or her choice with a minimum of verbal interaction. These problems are fundamentally different from traditional multiple choice tests in that:

1 the alternative solutions presented are exhaustive; there are no other alternative answers that are not nonsensical,

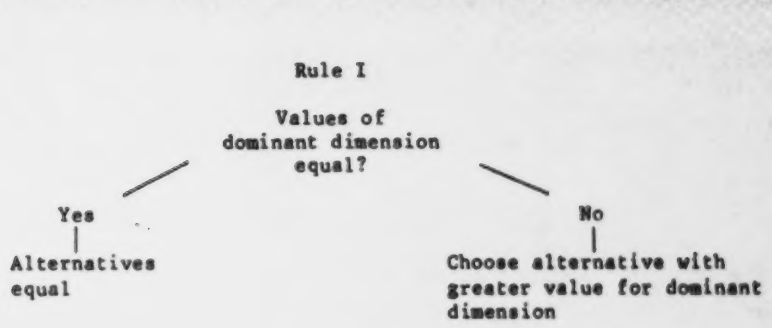


Figure 4.1 Siegler Rule I

2 the rules predict not only which problems a subject should answer correctly but also predict which problems will provoke guesses and which will be answered incorrectly (for this latter the rules specify which alternative will be chosen).

Siegler used the following description of the Piagetian Stages for the balance scale task as the basis of developing his rules:

> Stage 1
>
> the child understands that the weight is needed on both sides to achieve a balance and even that the weights should be approximately equal but there are as yet no systematic correspondences of the type 'further = heavier'. (Inhelder and Piaget, 1958, pp.168-169)
>
> Stage 2
>
> weight is equalized and added exactly, while distances are added and made symmetrical. But coordination between weight and distances as yet goes no further than intuitive regulation. (Inhelder and Piaget, 1958, p.169)
>
> Stage 3
>
> the subjects proceed from the same conception to a search for an explanation in the strict sense of the term . . . The general equilibrium schema is differentiated in the present case by constructing the proportions $W/W_1 = L_1/L$. (Inhelder and Piaget, 1958, pp.174-175)

Siegler calls the weight on each side of the fulcrum the dominant dimension because in cases of conflict, young children have been found to use weight more frequently than the distance of the weights from the fulcrum. He calls distance the subordinate dimension (Siegler, 1981, p.5). The postulated rules are:

> Rule 1 (see Figure 4.1). If the values of the dominant dimension are equal, then the alternative choices are equal. If not, then choose the alternative with the larger value for the dominant dimension.

Thus, the child using Rule 1 will not consider the distances of the weights from the fulcrum; to such a child, only the amounts of the weights matter. Stage 1 corresponds to Rule 1.

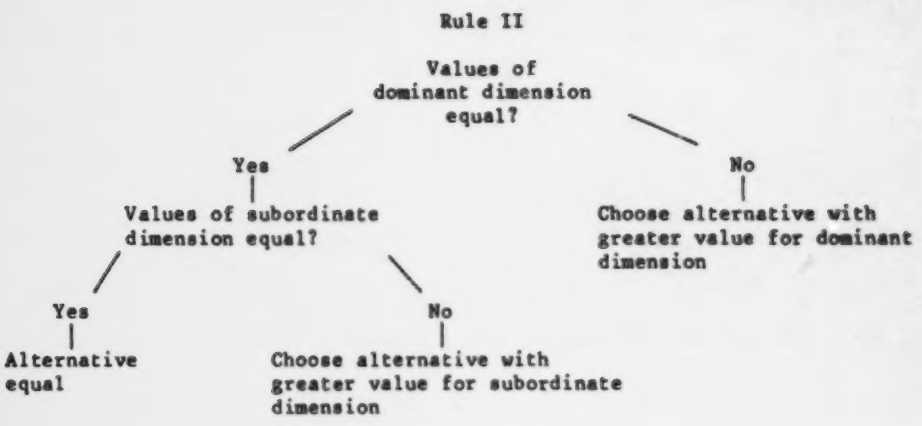


Figure 4.2 Siegler Rule II

Rule II (see Figure 4.2). If the values of the dominant dimension and the subordinate dimension are equal, then the alternative choices are equal. If the values of the dominant dimension are equal, but the values of the subordinate dimensions are not, then, choose the alternative with the larger value for the subordinate dimension. Otherwise, choose the alternative with the larger value for the dominant dimension.

A child using this rule will consider the distances of the weights from the fulcrum only when the weights are the same; otherwise this child will consider only the amounts of the weights.

Rule III (see Figure 4.3). Same as Rule II except that if the values of both of the dominant and subordinate dimensions are not equal, the child will 'muddle through' (Siegler, 1981, p.6).

A child using this rule is aware of his or her lack of understanding of the behaviour of the balance scale when both weights and distances vary, and will use some cognitive strategy such as guessing or taking cues from the experimenter. Rules II and III correspond to Stage 2.

Rule IV (see Figure 4.4). Use the correct formula for choosing the alternative (this will not necessarily involve actual calculation).

A child using this rule will compute torques on either side of the balance beam and choose accordingly. This computation may be either an actual calculation, or could be done 'by eye'.

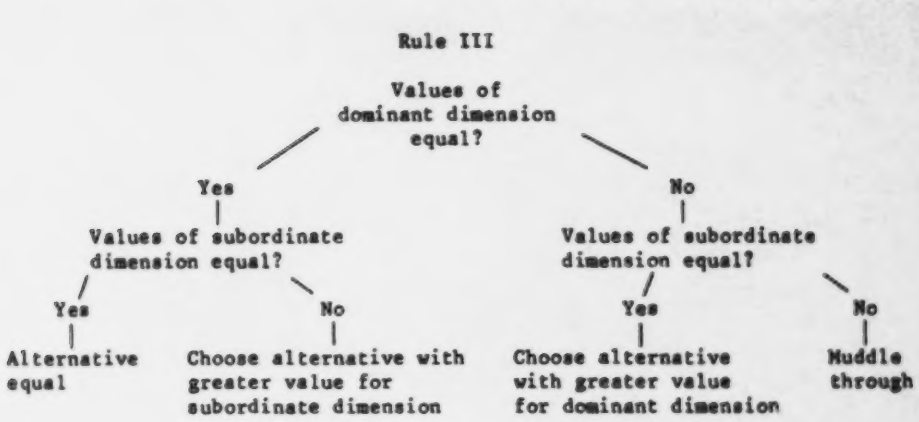


Figure 4.3 Siegler Rule III

In order to distinguish children at these four Rule levels, Siegler designed six problem types.

1 Equal Problems (E), with the same values on both dominant and subordinate dimensions for the two choices.

2 Dominant Problems (D), with unequal values on the dominant dimension and equal values on the subordinate dimension.

3 Subordinate Problems (S), with unequal values on the subordinate dimension and equal values on the dominant one.

4 Conflict-dominant Problems (CD), with one choice greater on the dominant dimension, the other choice greater on the subordinate dimension, and the one that is greater on the dominant dimension being the correct answer.

5 Conflict-subordinate Problems (CS), with one choice greater on the dominant dimension, the other choice greater on the subordinate dimension, and the one that is greater on the subordinate dimension being the correct answer.

6 Conflict-equal Problems (CE), with the usual conflict, and the two choices being equal on the outcome measure. (Siegler, 1981, p.9)

In the balance scale task the E problems would have both sides of the scale identical; with the D problems the distances would be the same, but the weights would vary and in the S problems the weights would be the same but the distances would vary. On the CD problems, the side with more weight will descend, on the CS problems, the side with the weight further from the fulcrum will descend, and on the CE problems the two sides balance, but both weights and distances are unequal.

The predicted success rates for each of these problem types for children using the four rules are given in Table 4.1. The six problem types give different profiles for the four rules, and this was the basis of Siegler's classification. The four rules do not, however, distinguish all of the item types: E and D are predicted to elicit identical

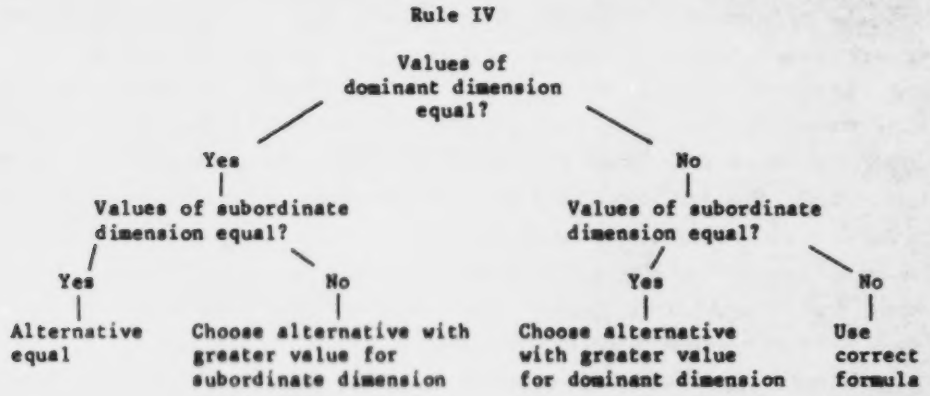


Figure 4.4 Siegler Rule IV

responses from children at all levels, as are CS and CE. Problem type CD has a distinctive predicted response pattern, indicating that children of a higher rule level (III) will give a lower rate of correct answers than will children at lower rule levels (I and II). This pattern causes no problem when using a data analysis technique that examines each problem type separately and uses complex rules to achieve a 'sensible' classification, as does Siegler. When using a probabilistic model such as Saltus, however, one assumes that a higher score is always modelled as indicating a higher ability (i.e. Rule) level. Hence, this problem type had to be left-out of all Saltus analyses. The monotonic relationship between problem types and rule levels that results when problem type CD is left-out, can be seen by considering the mean predicted success rates for problem types and rule levels as given in Table 4.1. These mean predicted success rates also illustrate the lack of predictive distinction between problem types E and D and between problem types CS and CE.

Table 4.1 Siegler Predictions for the Balance Scale Task

Problem type	Rule				Mean
	I	II	III	IV	
E	1.00	1.00	1.00	1.00	1.00
D	1.00	1.00	1.00	1.00	1.00
S	0.00	1.00	1.00	1.00	0.75
CD	1.00	1.00	0.33	1.00	0.83
CS	0.00	0.00	0.33	1.00	0.33
CE	0.00	0.00	0.33	1.00	0.33
Mean without CD	0.20	0.60	0.73	1.00	

The participants in the study were sixty subjects, ten at each of the following ages: 3 years, 4 years, 5 years, 8 years, 12 years, 21 years (college students). Half of each age group was male and the other half was female. The three tasks were administered twice, one month apart. For each task, the subjects were shown apparatus arranged to represent the problem types and asked to predict a certain result: for the balance scale, they had to predict whether the beam would dip left or right, or stay even. Each problem type was represented by 4 items, and, in order to be classified as belonging to a Rule level, the subject needed to answer 20 out of the 24 items in the way Siegler predicted. Additional sub-rules were used for some subjects at certain stages to check that the classification was accurate (Siegler, 1981, p.18).

Overall, Sielger found that the behaviour of most subjects 'fit' the Rule hierarchy well: for the four older groups, 96 per cent were classifiable in the balance scale task, 94 per cent in the shadows task and 79 per cent in the probability task (Siegler, 1981, p.22). The 3-year-olds, however, were found to give patterns of responses which predominantly resulted in no rule classification, and gave generally unrevealing explanations (Siegler, 1981, pp.23-26). These subjects were left out of the Saltus analyses because of this lack of consistency in their responses to the stimuli.

Siegler concluded that, for the balance scale, 'children were found to pass through a consistent age related sequence' (Siegler, 1981, p.26) and that, 'the developmental sequence on the projection of shadows task was very similar' (Siegerl, 1981, p.26), but that for the probability task, only Rules 1 and 4 were used with any regularity (Siegler, 1981, pp.26-27).

The Balance Scale Task

The balance scale rules and problems have been described in the previous section (see especially Figures 4.1 to 4.4 and Table 4.1). After removing the 3-year-olds, fifty subjects remained. One would not expect to find a great many stage transitions in the month between the two testings. In fact, Siegler found that, over the three tasks, 77 per cent of the subjects remained at the same rule level, 18 per cent advanced and 5 per cent moved to a lower level. For fifty subjects, this means that the net upward movement averaged across the tasks was 6.5 persons. This is not a large enough group to make a study of gains worthwhile, so the two testings will be treated as replications, to aid in the search for consistency. For the balance scale task, the first testing gave an intractable data set on the step from D to S. As the first testing also gave some intractable data sets for the shadows task, the analyses in this and the next section will use the second testing as the primary data set, and the results of the first testing will be discussed only when they provide interesting evidence for or against consistency.

Table 4.2 Rasch Results for Balance Scale

Item type	Mean difficulty	Standard error	Mean fit
E	-2.87	0.27	0.20
D	-2.16	0.25	0.11
S	0.00	0.20	-3.61
CS	1.73	0.19	2.49
CE	2.50	0.22	-0.56

An initial Rasch analysis of the five problem types, using the fifty subjects, gave the mean item difficulties and mean fit statistics in Table 4.2. The difficulties shown are centered on the average of the S items for consistency with later Saltus scales. The standard errors are calculated as the mean of the estimated standard errors of the problems, divided by the square root of the number of problems. The general pattern of the item difficulties conforms to the pattern predicted in Table 4.1: D and E are the easiest, S is in the middle and CS and CE are the hardest. However, the means for D and E are 1.93 standard errors apart, and those for CE and CS are 2.65 standard errors apart, which are considerably more than would be expected if the two pairs were actually equal in difficulty, as predicted. The fit statistics for the S problems show the large negative values that we expect to occur above a gap, which leads one to expect large asymmetry indices for the D to S and E to S steps. The total fit for CS is a large positive value which indicates a pattern of misfit different from that caused by rigidity. The total fit for CE is negative, but not large enough for one to be confident that it is indicating a gap. Thus, the Rasch results predict a gap for the D to S and E to S steps, hint that there might be a small gap from S to CE and imply a disorderly pattern for the step from S to CS.

Each Saltus analysis concentrates on just those parts of the sample that give information about each step; subjects who get all incorrect or all correct on both types of problems are not used by the procedure, since these subjects do not give any information concerning the relative difficulty of the problem types. With fifty subjects and what appear to be three levels of problem difficulty, this means that the number of subjects who give useful information regarding each of the two steps will be small. The numbers available for each of the Saltus analyses are shown in Table 4.3. The size of these samples makes the parameter estimates from the Saltus analyses have large standard errors. Nevertheless, many of the effects were large enough and consistent enough to warrant a detailed analysis of these data. Researchers considering collecting data for the investigation of developmental hierarchies should be aware that if n is the number of cases thought to be needed to make an analysis worthwhile, then n cases must be collected for <u>each</u> step.

Table 4.3 Number of Subjects in the Balance Scale Analyses

Step	First testing		Second testing	
	Subjects	Std. bias >2.0	Subjects	Std. bias >2.0
E to S	34	0	23	2
D to S	29	-	19	2
S to CS	36	5	34	4
S to CE	29	1	32	1

For the step from problem type E to problem type S, Saltus estimated a gap matrix of

$$\begin{bmatrix} 0.54 & 1.34 \\ -3.48 & 0.61 \end{bmatrix} \text{ with standard errors } \begin{bmatrix} 0.37 & 0.80 \\ 0.52 & 0.62 \end{bmatrix}$$

As this is in the range of gap matrices found to need correction for under-estimation of the BI gap parameter, simulations were carried out to correct for 'cross-overs' (Wilson, 1984, pp.114-125). The estimates of the group I gap indices from these tailored simulations are shown in Table 4.4 and illustrated in Figure 4.5. These simulations supported a linear correction which gave a corrected value of -4.16 for the BI gap parameter. This correction is large, and the values are not so far out on the logit scale

Table 4.4 Simulations for the E to S Step

Generator	Estimates	Mean of estimates
4.02	2.52	3.39
	3.21	
	3.49	
	3.66	
	4.55	
4.45	3.36	3.61
	3.46	
	3.51	
	3.62	
	4.09	
4.72	2.96	4.11
	3.82	
	4.44	
	4.59	
	4.75	
3.71	3.71	4.46
	4.26	
	4.34	
	4.60	
	5.38	

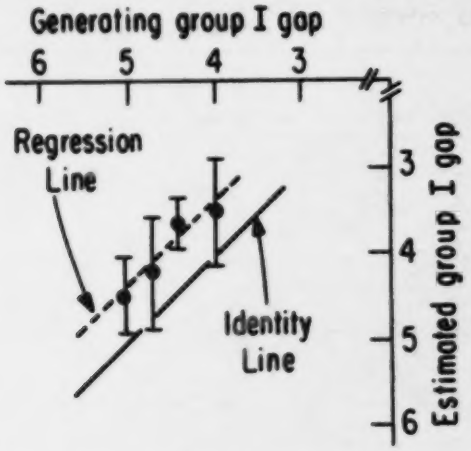


Figure 4.5 Group I Gaps for the E to S Simulations

that the correction has negligible practical impact. With this correction, the asymmetry index is 3.97 and its standard error is 1.20. The corrected item estimates for the E to S step are given in Table 4.5, the parameter estimates are given in Table 4.6 and the logit scale is illustrated in Figure 4.6.

These tables and the figure show a pattern similar to that for the subtraction tasks in Chapter 3. The problem types are quite homogeneous in difficulty and, for the group I subjects, the scale is strongly segmented, with average group I subjects succeeding quite well on average type E problems (probability of success = 0.63), but finding the S problems very difficult (probability of success = 0.02). The group II subjects find the E problems about equally as difficult as the group I subjects. They find the S problems somewhat harder than the E items, but not at all so hard as the group I students found them to be, and, in fact the problem types are not segmented for the group II students.

Table 4.5 Item Estimates for the E to S Step

Item type	Item	Group I Difficulty	Group I Standard error	Group II Difficulty	Group II Standard error
E	E1	-0.32	0.77	Same	
	E2	-0.32	0.77	as	
	E3	0.06	0.65	for	
	E4	0.58	0.54	Group I	
S	S1	4.08	1.02	0.71	1.29
	S2	5.21	1.08	1.24	1.34
	S3	4.68	1.08	1.71	1.29
	S4	4.24	0.96	0.27	1.29

Table 4.6 Score Estimates for the E to S Step

Person group	Score	Number of persons	Ability	Standard error
I	1	0	-1.04	1.22
	2	1	0.05	1.05
	3	4	1.02	1.07
	4	14	2.14	1.14
II	5	1	0.63	1.09
	6	0	1.25	1.16
	7	3	2.13	1.34

This pattern can be interpreted to mean that, for a subject working at the level of problem type E, the S problems are almost impossible, but for one working at the level of problems type S, the problem types are of roughly the same difficulty, with problem type S somewhat harder on the average. This can be compared to Siegler's predictions by noting that he would classify those who scored less than three as being below Rule I (at Rule 0), those who scored 3 or 4 out of the E problems as having attained Rule I, and those who scored at least 3 on both E and S as having attained Rule II. Thus, a Saltus score of 0 to 2 places the subject at Rule 0, a score of 3 to 5 places the subject at Rule I, and a score of 6 or 7 places the subject at Rule II (assigning the transitions between Rules to the lowest score possible).

Table 4.7 and Figure 4.7 compare the Saltus group classification with the rule-assessment classification and give both Saltus estimates and Siegler predictions of probability of success for a subject at each score, for the type E and type S problems. The two probability patterns are most discrepant at scores 5 and 6, where the Saltus transition is out of step with the Rule transition. They are more alike for scores 3 and 4 and for 7 and 8, where both classifications agree. Siegler does not make any prediction

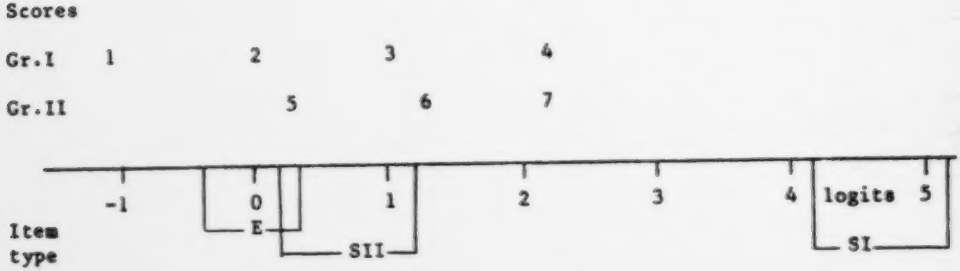


Figure 4.6 Logit Scale for the Balance Scale E to S Step

Table 4.7 Siegler and Saltus Classifications for the E to S Step

			Probability of success on average problem type			
			E		S	
Saltus group	Test score	Siegler rule	Saltus estimate	Siegler prediction	Saltus estimate	Siegler prediction
<II	0	0	0.26	?	0.00	0.00
I	1	0	0.26	?	0.00	0.00
I	2	0	0.51	?	0.01	0.00
I	3	I	0.73	1.00	0.02	0.00
I	4	I	0.89	1.00	0.07	0.00
II	5	I	0.65	1.00	0.48	0.00
II	6	II	0.78	1.00	0.63	1.00
II	7	II	0.89	1.00	0.80	1.00
>I	8	II	0.89	1.00	0.80	1.00

for the success on problems type E of those who score less than 3, but his prediction of success for these subjects for problem type S agrees well with the Saltus estimates. It should be remembered when interpreting this table and this figure that both Saltus estimates and Siegler's predictions are not applicable to students who do not fit (according to their respective patterns of misfit). In considering the probability patterns in Figure 4.7, it must be recalled that Siegler's predictions represent an ideal; that he does not expect that ideal to be attained is indicated by his acceptance of 3 out of 4 as sufficient evidence that a rule has been achieved. So the position of the probability curve is not as important as its shape. For problem type E, Siegler predicts that all subjects with scores from 3 to 7 will have equal chance of success, and that is the pattern that Saltus gives. For problem type S, Siegler predicts a large increase at score 5, and Saltus finds a large increase at 4; this is not a crucial difference, the important point is that both agree on the existence of a jump.

The small number of subjects bars detailed interpretation of the Saltus results. (A movement of one standard error below or above the logit location of a score of 5 results in a range of success for problem type E from 0.39 to 0.85 and for problem type S from 0.23 to 0.73.) But, in their general tendencies, the two patterns of probability are similar:

1 Both show a fairly constant success rate on problem type E for persons at score 3 and above, but Saltus gives a lower (and more realistic) rate.

2 Both show a dramatic increase in rate of success for problem type S for subjects above a cut-off score, although they differ, by one score point, on where that cut-off is.

Although one must be guarded in making conclusions, one can say that the estimated Saltus model for the E to S step matches the Siegler's predictions for the rule-assessment

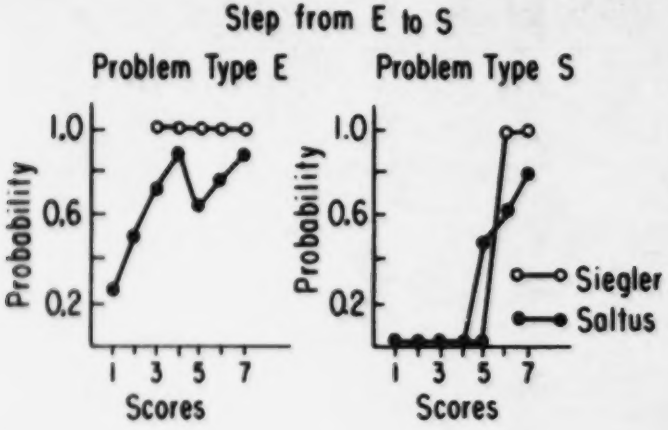


Figure 4.7 Siegler Compared to Saltus for the E to S Step

model. As there were only two large misfits in the Saltus analysis (see Table 4.3), the Saltus result can be considered a confirmation that the great majority of subjects are following the rule-assessment pattern on this step.

The step from problem type D to problem type S in the second testing gave an asymmetry index of 3.17 (standard error = 1.48) and a similar gap matrix,

$$\begin{bmatrix} 0.64 & 1.48 \\ -4.01 & 0.52 \end{bmatrix} \text{ with standard errors } \begin{bmatrix} 0.45 & 1.06 \\ 0.60 & 0.70 \end{bmatrix}$$

and only two misfits, so the pattern is constant, as Siegler predicted, across problem types E and D. The BI gap matrix for this step was adjusted by a series of simulations similar to those for the E to S step. The estimates from these simulations are given in Table 4.8 and illustrated in Figure 4.8; they indicate that a linear correction of the original group I gap index from 4.39 to 4.65 was suitable. Thus for the transition from Rule I to Rule II, the results of the Saltus analyses agree with Siegler's conclusions. The segmentation of the problem types and the large positive asymmetries indicate that the step from Rule I to Rule II, as realized by problem types E, D and S, could be a step in a hierarchical development.

The Rule II to Rules III and IV transitions were examined by two Saltus analyses: one for the step from problem type S to problem type CS and one for the step from problem type S to problem type CE. The asymmetry index for the S to CS step for the second testing was -1.36 with a standard error of 0.70 and the gap matrix was

$$\begin{bmatrix} -0.23 & 2.52 \\ -1.24 & 0.15 \end{bmatrix} \text{ with standard errors } \begin{bmatrix} 0.27 & 0.52 \\ 0.29 & 0.25 \end{bmatrix}$$

Table 4.8 Simulations for D to S Step

Generator	Estimates	Mean of Estimates
4.39	3.01	4.05
	3.36	
	3.87	
	4.83	
	5.20	
4.64	3.79	4.23
	3.91	
	3.91	
	4.64	
	4.89	
4.84	4.16	4.11
	4.54	
	4.57	
	5.01	
	5.92	
5.05	3.71	4.94
	4.39	
	4.96	
	5.38	
	6.25	

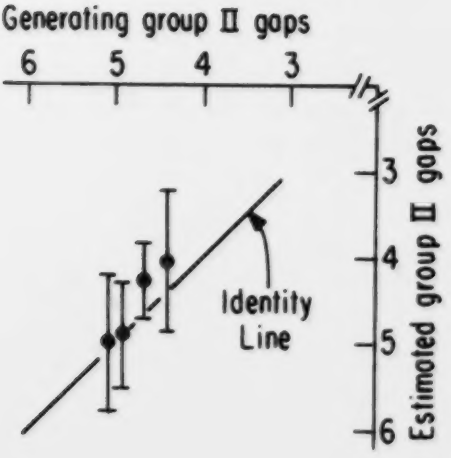


Figure 4.8 Group I Gap for the D to S Step

Table 4.9 Item Estimates for the S to CS Step

Item type	Item	Group I Difficulty	Group I Standard error	Group II Difficulty	Group II Standard error
S	S1	0.00	0.48	Same	
	S2	0.17	0.47	as	
	S3	0.00	0.48	for	
	S4	-0.18	0.51	Group I	
CS	CS1	1.19	0.55	2.55	0.69
	CS2	0.75	0.54	2.13	0.68
	CS3	0.75	0.54	2.13	0.68
	CS4	1.30	0.56	2.66	0.69

The item estimates for step S to CS are given in Table 4.9, the score estimates in Table 4.10 and the logit scale is illustrated in Figure 4.9. The subjects scoring 4 and below are referred to as group II and those scoring 5 and above are referred to as group III. This step presents a different picture from that for the previous step. The score estimates do not fold-back in the pattern that has been observed at other stage transitions, and the asymmetry index is negative: -1.36. This means that the CS items for group I occur to the left of the CS items for group II on the logit scale, so that there are subjects in group I (those with score 4) who are predicted to do better on the CS items than some in group II (those who score 5). Although the Saltus analysis is capable of modelling this situation, it has not been discussed to this point because it does not conform to the rigidity pattern for a hierarchical theory.

To understand this result, consider the probability of success that Saltus gives for different scores, as shown in Table 4.11 and Figure 4.10. The probabilities for problem type S show a steady increase as the score increases; there is no plateau as there was for type E in the previous step (see Table 4.7 and Figure 4.8). The probabilities for the CS problems do not show the characteristic jump at the boundary between groups (as for

Table 4.10 Score Estimates for the S to CS Step

Person group	Score	Number of persons	Ability	Standard error
II	1	1	-1.34	1.11
	2	1	-0.46	0.88
	3	1	0.16	0.80
	4	12	0.71	0.78
III	5	9	1.69	1.00
	6	6	2.44	1.05
	7	4	3.51	1.23

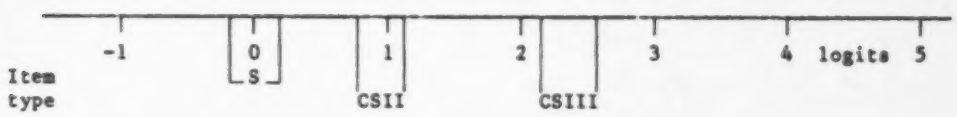


Figure 4.9 Logit Scale for the Balance Scale S to CS Step

problem type S in the previous step), but instead they flatten out across the boundary and actually decrease from score 4 to score 5.

Before attempting to interpret this strange result, I made some investigations to rid myself of doubts over its reliability. First, there were four subjects in the analysis who gave large standardized biases; they did much better on the CS problems than on the S problems, and moreover, these same four appeared as misfits in the first testing for exactly the same reason. These four having been identified as consistent misfits, they were deleted from the data set and the Saltus analysis was run again. The resulting gap matrix was

$$\begin{bmatrix} 0.25 & 3.65 \\ -2.38 & 0.04 \end{bmatrix} \text{ with standard errors } \begin{bmatrix} 0.37 & 1.01 \\ 0.45 & 0.26 \end{bmatrix}$$

Table 4.11 Siegler Predictions and Saltus Estimates of Success for the S to CS Step

			Probability of success on average problem type			
			S		CS	
Saltus group	Test score	Siegler rule	Saltus estimate	Siegler prediction	Saltus estimate	Siegler prediction
<III	0	II	0.21	0.00	0.09	0.00
II	1	II	0.21	0.00	0.09	0.00
II	2	II	0.39	0.00	0.19	0.00
II	3	II	0.54	1.00	0.30	0.00
II	4	II	0.67	1.00	0.43	0.00
		III	0.67	1.00	0.43	0.33
III	5	III	0.84	1.00	0.34	0.33
III	6	IV	0.92	1.00	0.52	1.00
III	7	IV	0.97	1.00	0.76	1.00
>II	8	IV	0.97	1.00	0.76	1.00

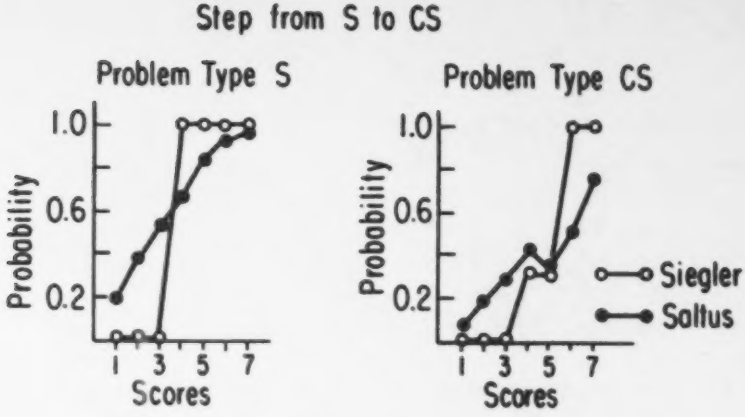


Figure 4.10 Siegler Compared to Saltus for the S to CS Step

The asymmetry index is still negative, but a little smaller than for the full data set: -0.98 with standard error 1.19. Second, a series of five simulations was run to check on the bias of the gap estimates. The mean group II and group III gap indices were 1.54 with a standard error of 0.17 and 1.79 with a standard error of 0.19. As the generators were 1.47 and 2.37, this suggests that the group II gap index was underestimated; that is, is true value is higher, making the asymmetry index even more negative. These investigations suggest that the phenomenon is probably not a product of random fluctuation in the data.

A clue to the possible meaning of the plateau in the CS probabilities in Table 4.10 is provided by Siegler's predicted probabilities. First, however, the rationale, following Siegler, for classifying the scores into Rules (as shown in the third column of Table 4.11) must be given. As before, a subject is classified below Rule II until he or she gets 3 out of 4 of the S problems correct; a score of 0 to 2 corresponds to being below Rule II, and a score of 3 indicates that Rule II has been attained. Here the classification differs from that for the previous step, because the predicted jump in probability is from 0.00 to 0.33 rather from 0.00 to 1.00 (compare the last columns of Tables 4.7 and 4.11). This means that, once at rule II, a subject need only demonstrate success on one-third of the four problems in CS, or 1.33 problems, to be considered as having attained Rule III. Unfortunately, 1.33 is not an integer, so some interpretation is needed. I decided that if a subject was correct on at least 1 out of 4, then Rule III had been achieved. When a subject gets 3 out of 4 on both problem types, Rule IV has been attained, so scores 6 to 8 are assigned to Rule IV.

This results in the pattern of probabilities shown in columns 5 and 7 of Table 4.10. The interesting thing to note in this Table is that the plateau of probabilities for the CS problems occurs just where Siegler predicts that subjects will be guessing. Given that Siegler has succeeded in designing items that provoke guessing at certain rule levels, and if these rules are indeed determining the behaviour of these subjects, then guessing could explain why the CS items are not discriminating between subjects with score 3, 4 and 5. Thus, the reversal of the usual increase in probability of success, which Saltus has indicated for score 5, matches Siegler's prediction. In general, a negative asymmetry index will cause a similar reversal, or at least a plateau, in the estimated probabilities of success. The causes will not always be the same, but it seems likely that Siegler is correct here and that guessing is the origin of the negative asymmetry index.

The Siegler prediction and the Saltus estimates for problem type S, as given in Table 4.11 and Figure 4.10 are not so well-matched as those for problem type CS. Siegler predicts a large jump in the probabilities near score 3, but this is not observed in the Saltus estimates. It may be that the guessing which caused the plateau in the CS items has caused the probability jump for the S problems to be obscured.

Overall, the match between Siegler's predictions and the Saltus estimates is good for the CS problems, but not conclusive for the S problems. The existence of four subjects who consistently found the CS items easier than the S items is of concern, and, were they still available, could lead to interesting follow-up study. Although Saltus has shown that it can model this guessing, the use of items which elicit guessing is not a sound procedure in the investigation of hierarchies. The guessing pattern (as indicated by a negative asymmetry index) has overwhelmed any chance of finding evidence of rigidity (which would be indicated by a positive asymmetry index).

The Saltus analysis for the second testing for the step from S to CE gave an asymmetry index of 0.99 with a standard error of 1.54, and the following gap matrix:

$$\begin{bmatrix} 0.54 & 3.87 \\ -4.26 & 0.06 \end{bmatrix} \text{ with standard errors } \begin{bmatrix} 0.40 & 1.01 \\ 1.05 & 0.28 \end{bmatrix}$$

The item estimates are given in Table 4.12, the score estimates in Table 4.13, and the logit scale is illustrated in Figure 4.11. The problem types segment the logit scale, but the CE problems are not homogeneous in difficulty and the overlap between the location of the CE problems for groups I and II is considerable.

The difference between the mean location of the CE items for group II and group III, 0.93, is not large compared to the standard errors for the two means, 0.62 and 0.60. This lack of distinction between the locations for the two person groups, combined with the spread in difficulty for the CE items, has resulted in a pattern of probabilities, as shown in Table 4.14 and Figure 4.12, that does not clearly exhibit the large jumps expected of hierarchical situations. Once again, the Saltus probability pattern for the S

Table 4.12 Saltus Item Estimates for the S to CE Step

Item type	Item	Group II Difficulty	Group II Standard error	Group III Difficulty	Group III Standard error
S	S1	0.01	0.74	Same as for Group I	
	S2	0.40	0.70		
	S3	0.01	0.74		
	S4	-0.42	0.78		
CE	CE1	4.27	1.22	3.28	1.16
	CE2	5.57	1.24	4.58	1.18
	CE3	6.06	1.27	5.07	1.21
	CE4	3.30	1.24	2.31	1.81

problems is not conclusively like or unlike the pattern for the rule-assessment model; there are large increases in probability, not a single large jump, as predicted, but the predictions and the estimate match well for the higher scores. The patterns for problem type CE are not conclusive either: the plateau which was associated with guessing for the S to CS step is not apparent in these results; but the predictions and the estimates match well for the lower scores. It seems that the CE items are not provoking guessing to the same extent as were the CS items. The Saltus analysis for this step does not conclusively support Siegler's claim that the rule-assessment predictions are borne out in the data, nor conclusively refute that claim.

The preceding discussion on the S to CS and S to CE steps used only three groups, but the rule-assessment theory was designed with four rules. This is no oversight. The four rules were to be distinguished by five problem types, but the predicted responses given in Table 4.1 indicate that the rules differ on only three sets of the problem types. Thus Saltus, because it demands a matching of subject group with the problem type that characterises that group, could only distinguish three groups of subjects. Nevertheless, the Saltus analysis has shown support for Siegler's claim that the subjects are performing

Table 4.13 Score Estimates for the S to CE Step

Person group	Score	Number of persons	Ability	Standard error
II	1	4	-1.09	1.22
	2	1	0.00	1.07
	3	0	1.00	1.11
	4	6	2.25	1.25
III	5	8	2.68	1.46
	6	8	3.82	1.48
	7	5	5.11	1.60

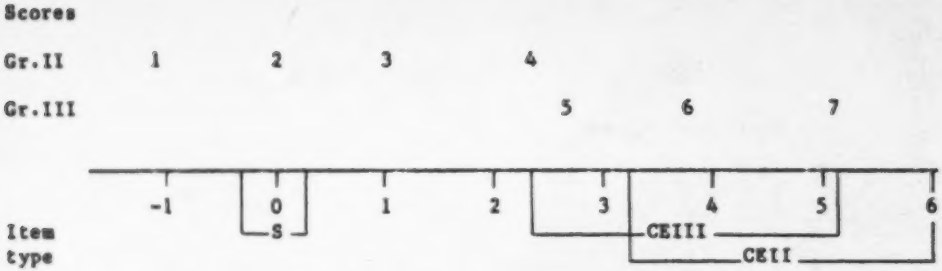


Figure 4.11 Logit Scale for the Balance Scale S to CE Step

according to the prediction of the rule-assessment model for the step from Rule I to Rule II. But the Saltus analyses have not clearly supported this claim for the remaining two steps although some good matches between prediction and estimate were found for some problem types. The analyses are consistent with the hypothesis that the problem may be due to the incitement to guess present in the CS problem type, and in part due to the lack of homogeneity of the CE problems. With respect to a generic theory of hierarchical development, the analyses have indicated: that the E and D to S steps behave like steps in a hierarchy; that the S to CE step, although it has a sizeable positive asymmetry index, does not behave hierarchically, perhaps because of the lack of homogeneity of the CE items; and that, although the S to CS step exhibits segmentation, guessing has obscured the potential for demonstrating the hierarchical nature of this step.

Table 4.14 Siegler Predictions and Saltus Estimates for the S to CE Step

			Probability of success on average problem type			
			S		CE	
Saltus group	Test score	Siegler rule	Saltus estimate	Siegler prediction	Saltus estimate	Siegler prediction
<III	0	II	0.25	0.00	0.00	0.00
II	1	II	0.25	0.00	0.00	0.00
II	2	II	0.50	0.00	0.01	0.00
II	3	II	0.73	1.00	0.02	0.00
II	4	II	0.90	1.00	0.07	0.00
		III	0.90	1.00	0.07	0.33
III	5	III	0.94	1.00	0.24	0.33
III	6	IV	0.98	1.00	0.50	1.00
III	7	IV	0.99	1.00	0.79	1.00
>II	8	IV	0.99	1.00	0.79	1.00

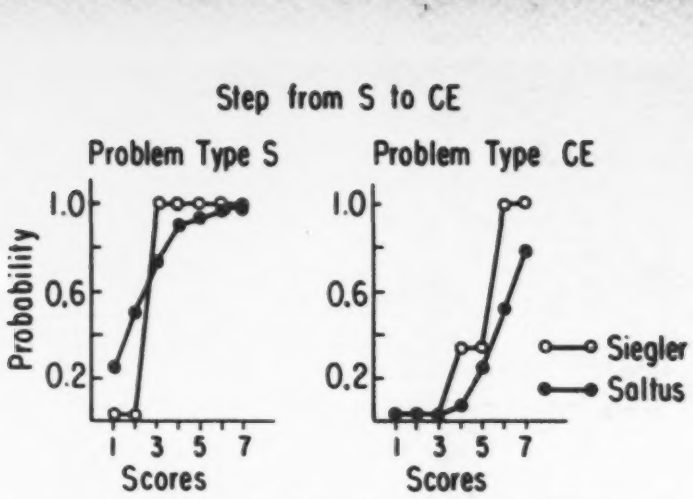


Figure 4.12 Siegler Compared to Saltus for the S to CE Step

Linking the Saltus Analyses

The three Saltus analyses for the balance scale task described in detail above are shown linked together on the one logit scale in Figure 4.13. For the balance scale tasks, the asymmetry indices are non-zero, so the BI and BII locations are far apart on the logit scale. The method of linking is determined by two considerations:

1 The S problem type is in every step, so it can be the basis for the linking procedure within tasks.

2 The location of the group I students will be used later to link all three tasks, so the location of the S problem type for these students is used to link the other problem types.

Thus, for each of the analyses, the logit scales were translated to make zero coincide with the mean difficulty of the S items. The estimates for each subject group were separated, so that the two sets of analyses, D and E to S and S to CS and CE, become three distinct segments of the one scale, one segment for each of the person groups. The D to S and E to S steps supplied the difficulties for problem types D, E and S as they applied to group I, and these are given as the first row below the logit scale in Figure 4.13. The second row gives the locations for D, E and S for group II taken from steps D to S and E to S, and it also gives the locations of CS, CE and S for group II taken from steps S to CS and S to CE. The third row gives the location for CS, CE and S for group III, taken from steps S to CS and S to CE. The location given for the person groups deserves some comment. The mean of group I, for example, was located at -4.13 by the E to S step and -4.06 by the D to S step; the location given is the average of the two. The group II location is the average of 0.61 (E to S), 0.56 (D to S), -0.23 (S to CS) and 0.54 (S to CE). The group III location is the average of 2.52 (S to CS) and 3.87 (S to CE).

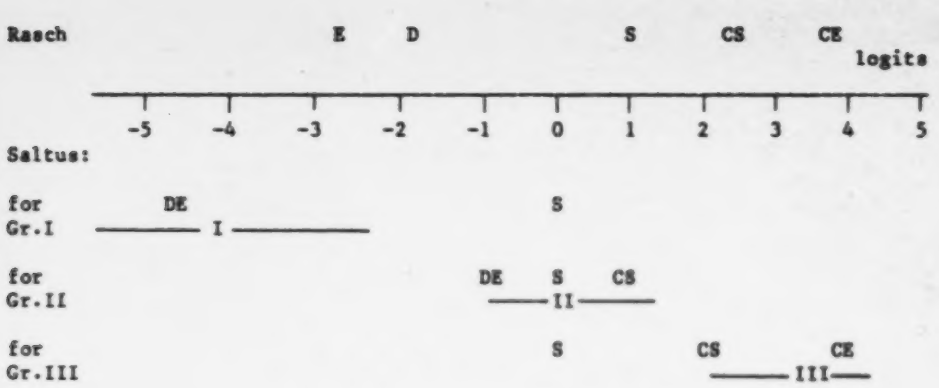


Figure 4.13 Linked Logit Scale for the Balance Scale Task

The lower limit for each ability group is similarly the average of the lowest scores from the appropriate steps, and the upper limit is the average of the highest scores from the appropriate steps. For comparison, the means of the Rasch estimates for each problem type are shown above the logit scale.

This figure illustrates the segmentation of the logit scale achieved by Siegler's problem types, and the differences in item difficulty for the different subject groups. The unoccupied regions between the segments of the logit scale, 1.74 logits between groups I and II, and 0.51 between groups II and III, show that, especially between groups I and II, Siegler's problems have succeeded in distinguishing between the groups very well. The differences in the location of the problem types for the person groups has been examined in the previous section and will not be repeated here, except to note that, for a subject at the top of group II, the CS problems are somewhat easier than for a subject at the bottom of group III, which is consistent with the interpretation that these items are provoking guessing in the less able subjects. Thus the linked scale provides another way of displaying the anomalous behaviour of the CS problems.

The gap matrices for the four Saltus analyses for the second testing are presented again in Table 4.15 beside the equivalent gap matrices for the first testing. (All estimates presented are in the original, uncorrected, form.) The entries in the gap matrices for the two testings are all within a standard error of one another. The pattern of asymmetry indices is also repeated. The E to S step gives a large positive asymmetry index, the S to CS step gives a small negative index and the S to CE step gives a small positive index. This stability gives one confidence that the patterns discussed above are not ephemeral, although not so much confidence as would be the case given a larger sample and (hence) smaller standard errors.

Table 4.15 Saltus Matrices from First and Second Testings of the Balance Scale Task

Step	First Testing				Second Testing			
	Gap Matrix		Standard Errors		Gap Matrix		Standard Errors	
E to S	0.48	1.63	0.39	0.43	0.54	1.34	0.37	0.80
	-2.83	0.51	0.43	0.29	-3.48	0.61	0.52	0.62
D to S	-	-		0.64	1.48	0.45	1.06	
					-3.75	0.52	0.60	0.70
S to CS	-0.02	2.30	0.22	0.60	-0.23	2.52	0.27	1.06
	-1.70	0.22	0.31	0.32	-1.24	0.15	0.29	0.25
S to CE	-0.41	2.30	0.32	0.60	0.54	3.87	0.40	1.01
	-3.09	0.24	0.61	0.30	-4.26	0.06	1.05	0.28

The problem types, rules and predictions of success for the projection of shadows task are the same as for the balance scale task (i.e. Table 4.1 applies to these problems as well as the balance scale problems). The apparatus for this task consisted of two lights projecting shadows on a screen from two horizontal cross-bars. The length of the cross-bars and their distances from the lights could be manipulated to give different problems; length is considered the dominant dimension and distance the subordinate (Siegler, 1981, pp.5-16). Subjects were asked to predict whether the shadow to the left or the right would be longer if the lights were turned on, or if they would be the same length. The gap matrices for the projection of shadows task are given in Table 4.16 and the linked logit scale is illustrated in Figure 4.14.

The figure shows a quite similar pattern to that for the balance scale task. The logit scale is well-segmented, the E and D problems appear closer in difficulty to the S problems for person group II than for person group II (corresponding to positive asymmetry indices for the E to S and D to S steps), the CS problems show the same anomalous behaviour, being relatively easier for some group I subjects than for some

Table 4.16 Saltus Statistics for the Second Testing of the Shadows Task

Step	Gap Matrix		Standard Errors	
E to S	0.83	3.15	0.40	1.04
	-4.52	0.19	0.72	0.42
D to S	0.55	3.13	0.37	1.02
	-3.40	0.08	0.47	0.40
S to CS	-0.34	3.54	0.19	1.01
	-1.26	0.08	0.24	0.35
S to CE	0.27	4.05	0.30	1.01
	-2.89	-0.02	0.51	0.33

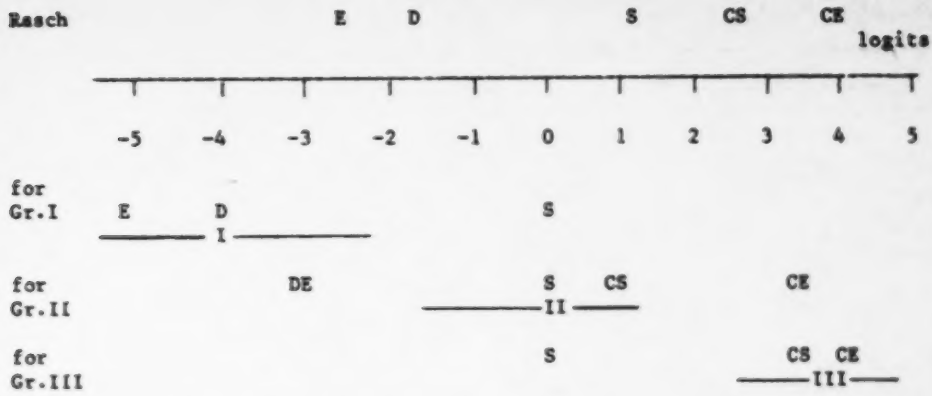


Figure 4.14 Linked Logit Scale for the Shadows Task

group II subjects. Thus, the same conclusions can be drawn for this task as for the previous task. The rule-assessment predictions are borne out quite well by the Saltus analyses for the steps D to S and E to S, and these steps follow a pattern in keeping with a developmental hierarchy. However, the upper two steps, although showing a progression in difficulty, follow neither the pattern predicted by Siegler, nor that for a developmental hierarchy.

The definition of problem types for the probability task was different from that for the other two, and so the predicted probabilities of success differ also. As Siegler gives only one example of each type and no explicit definition, the types will not be described here. Fortunately, for our purposes, the predictions of success are sufficient to define their natures. These predictions are given in Table 4.17. The arrangement is similar to that for the other two tasks, but there are now four different patterns of probability: A and C are the easiest types, E and F are the hardest, and B and D fall in between, with B being easier than D. The tendency to provoke guessing is not predicted for any of the problem types.

Table 4.17 Siegler Predictions for the Probability Task

Problem type	Rule			
	I	II	III	IV
A	1.00	1.00	1.00	1.00
B	1.00	1.00	1.00	1.00
C	0.00	1.00	1.00	1.00
D	0.00	0.00	1.00	1.00
E	0.00	0.00	0.00	1.00
F	0.00	0.00	0.00	1.00

Table 4.18 Saltus Statistics for the First Testing of the Probability Task

Step	Gap Matrix		Standard Errors	
A to B	0.20	1.05	0.29	0.39
	-2.23	0.82	0.35	0.36
C to B	0.29	0.59	0.33	0.25
	-2.56	1.34	0.38	0.31
B to D	-0.32	1.06	0.36	0.52
	-1.14	0.81	0.42	0.48
D to E	-0.27	1.62	0.36	0.49
	-1.22	-0.43	0.44	0.33
D to F	-0.20	2.34	0.30	0.73
	-1.33	0.20	0.37	0.30

The apparatus for this task consisted of two sets of marbles, with different numbers of red and blue marbles in each. The subjects were asked to choose which pile gave the better chance of picking a blue marble, if one had to pick a marble with eyes closed (Siegler, 1981, pp.15-16). The gap matrices for the probability task are given in Table 4.18 and the linked logit scale is illustrated in Figure 4.15.

The linking in Figure 4.15 is also somewhat different from that for the previous two tasks. The mean difficulty is used to position the estimates from the A to B, C to B, and B to D steps, but the mean difficulty of problem type D is used to position the estimates from the D to E and D to F steps, and the two are linked to one another through the B to D step. The pattern shown in this figure is quite different from that for the other two tasks. The most striking feature is that group I and II are not differentiated by the problem types. There is quite good segmentation between group I and group II, but the segmentation between groups III and IV is not very pronounced. The problem types show strong positive asymmetry, and hence, strong rigidity, for steps A to B and C to B, while the B to D and D to E steps show only little asymmetry, and the D to F step is of a similar type to the S to CS step in the balance scale task. The C to B step gives a negative group II gap index so that problem type C is easier than type B for the group II students. This swapping of problem difficulties will not be interpreted because of the high number of misfits in the C to B analysis (12 out of 37 subjects gave standardized biases over 2.0). Taken together, however, these two results imply considerable problems with the specification of problem type C. Given this difficulty with problem type C, the only step that can be said to clearly follow the prediction is the A to B step, the rest do not show the kind of behaviour associated with either the prediction of the rule-assessment model or the steps of a developmental hierarchy.

Summarizing his findings on developmental sequences between tasks, Siegler found that Rules I, II and III were acquired earlier on the balance scale task and the projection

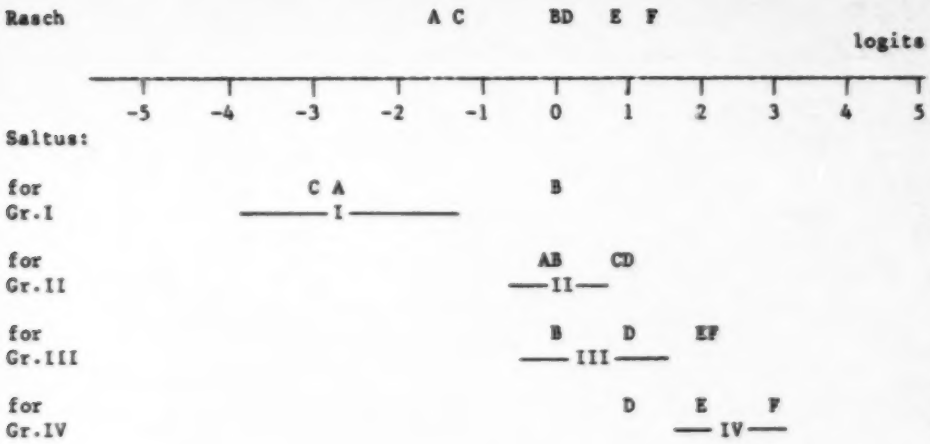


Figure 4.15 Linked Logit Scale for the Probability Task

of shadows task than on the probability task, but that Rule IV was acquired much earlier on the probability task than the other two. He also noted that, at the individual level, there was little synchrony between the tasks (Siegler, 1981, pp.27-28). In order to make the same type of comparisons using Saltus, it is necessary to link the three logit scales together. Common items were used to link the scales within tasks; as there are no common items between tasks, common subjects must be used to link across the tasks. Ten subjects were found to be common to all the group I to group II analyses, that is, for steps E to S and D to S for the balance scale and the projection of shadows tasks, and the A to B and C to B steps for the probability task. These subjects are listed in Table 4.19, and their locations on each of the logit scales as given by the Saltus analyses. The mean of the group for each analysis was calculated and then averaged for the two analyses per

Table 4.19 Common Subjects Used to Link the Three Logit Scales

	Balance		Shadows		Probability	
Student	E to S	D to S	E to S	D to S	A to B	C to B
31	-2.53	-2.38	-2.47	-1.84	-1.02	-1.30
32	-2.53	-2.38	-2.47	-1.84	-0.07	1.04
35	-2.53	-2.38	-2.47	-1.84	-0.07	0.17
37	-2.53	-2.38	-2.47	-1.84	-1.02	1.91
38	-2.53	-2.38	-2.47	-1.84	-2.52	-1.30
39	-2.53	-2.38	-4.06	-1.84	-1.02	0.17
40	-3.65	-2.38	-2.47	-1.84	-1.02	-1.30
41	-2.53	-2.38	-2.47	-1.84	-1.02	-1.30
44	-2.53	-2.38	-2.47	-1.84	-1.74	-2.09
45	-2.53	-4.59	-5.18	-3.88	-1.02	-2.09
Mean	-2.64	-2.60	-2.90	-2.04	-1.05	-0.74

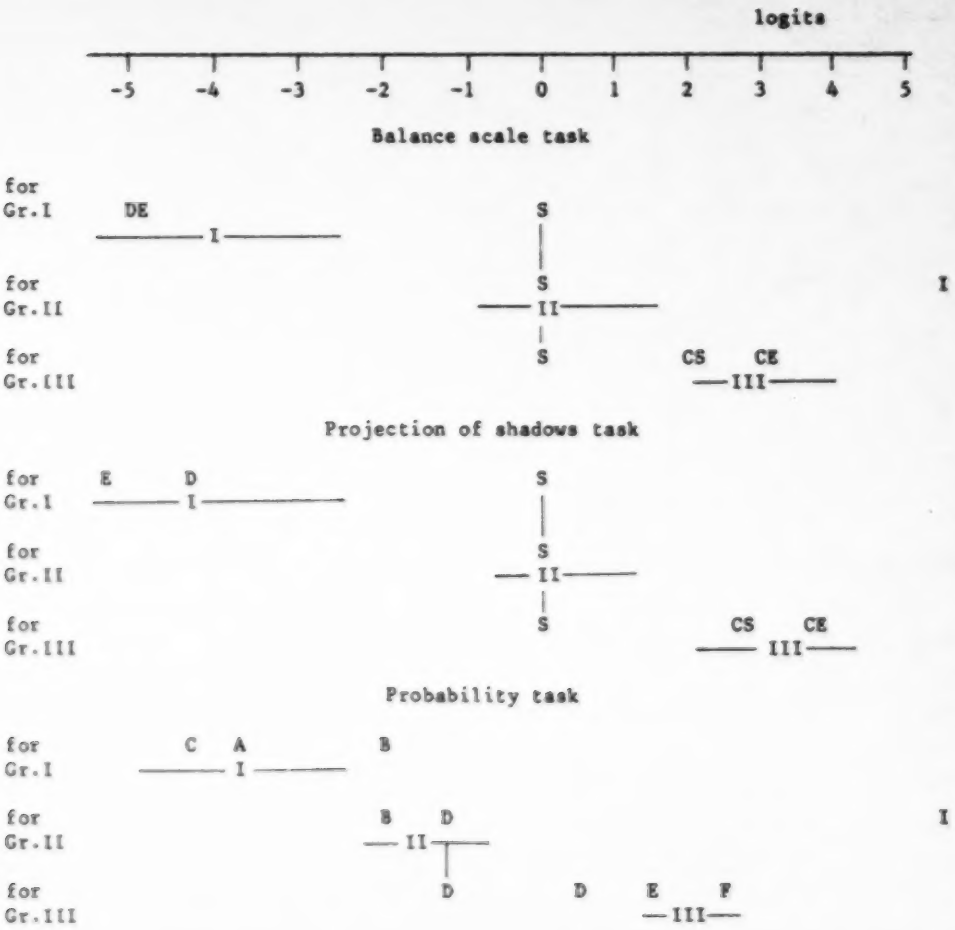


Figure 4.16 Linked Saltus Logit Scale for the Three Tasks

logit scale. This indicated that, if the balance scale estimates are taken as the standard, the locations on the projection of shadows scale need to be translated by -0.15 logits and those on the probability scale need to be translated -1.72 logits. The final, fully-linked, logit scale is illustrated in Figure 4.16. Note that the groups II and III for the probability task have been collapsed to group II and that group IV has been re-named group III. Three Rasch analyses for the three tasks are also displayed in Figure 4.17.

The situation portrayed in this figure agrees with Siegler's initial findings. The first two problem types and group I for the probability task occur somewhat to the right of those for the other two tasks; in other words, the first two problem types are more difficult, and those who have shown some success on them are more able, for the

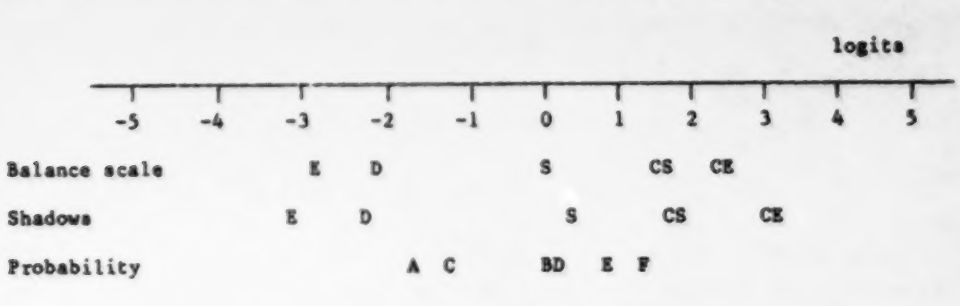


Figure 4.17 Linked Rasch Logit Scale for the Three Tasks

probability task than for the other two. And the last two problem types are easier, and those who are succeeding on them less able, for the probability task than for the other two, as is shown by the location of types E and F to the left of CS and CE and by the location of group III for the probability task to the left of the locations of group III for the other two tasks. However, contrary to Siegler's findings, the first two problem types and group I for the probability task occur at approximately the same location as those for the other two tasks; in other words, the first two problem types are no more difficult, and those who have shown some success on them are no more able, for the probability task than for the other two. The Rasch scale (Figure 4.17) agrees with Siegler's findings: the Rasch and Siegler results differ from those for Saltus because they are both averaging the position of the A and C problem types for all the subjects, whereas Saltus is using the responses of only those who are actually learning types A and C - the group I students. Despite this difference between the probability task and the other two, Figure 4.16 (the Saltus scale) shows considerable synchrony in the placement of the subject groups. The three group I's occupy a different region of the logit scale from the three group II's, and, except for the probability task, the group II's occupy a different region from the group III's. The problem types are not so well-behaved, but the defining problem types for each group - E and D (from the first two tasks) and C and A for group I, S and B and D (from the probability task) for group II, and CS and CE and E and F for group III - all fall within distinct regions of the logit scale. Although Siegler found little synchrony at the individual level, there is agreement among the tasks at the group level, if the groups chosen to display this are the three Saltus groups rather than Siegler's four rule groups. As Rules II and III correspond to Piaget's Stage 2, this pattern suggests that the Piagetian classification into three stages is better represented by these data than the four Siegler rule levels.

CHAPTER 5

CONCLUSION

Background to the Saltus Model

The Saltus model for the analysis of data from hierarchical theories of development originated in the need for a psychometric model that articulated psychological and educational theories like those of Piaget (1960) and Gagné (1962, 1968). The insight contained in their theories which is embodied in Saltus, is that learning is a process of growth, akin to biological growth; it can occur smoothly, or in spurts or stages, and once it has occurred it cannot be undone. This learning process has been termed development, and when it occurs in stages, hierarchical development. Piaget has offered his stages as a way of representing hierarchical development; Gagné, learning hierarchies. The features of these schemes that are modelled by Saltus are summarized by the twin concepts of gappiness and rigidity. Gappiness is the lack of a stable state between stages of development. Rigidity is the fixity of the sequence of stages. The theories of Piaget and Gagné contain many more ideas than are expressed by gappiness and rigidity; Saltus attempts to model only those features of them that make development hierarchical. The Rasch model (1960/1980) is a probabilistic method for analysing test data that provides a clear and interpretible scale of person ability and item difficulty that is well suited to the interpretation of development through stages of a hierarchy. There is, however, no explicit way to integrate the knowledge of the stage origin of items into the model (although, through the use of fit statistics, some clues can be gained). Saltus is an attempt to adapt the Rasch model to the problem of developmental hierarchies, while maintaining the advantages of the Rasch approach to measurement; the method of adapting the Rasch model was suggested by the Linear Logistic Test Model. As a developmental hierarchy is composed not merely of distinct types of items, but also of groups of persons who behave differently when attempting the different item types, the item and person parameters for Saltus are separated into those that operate within each person group and item type and those that operate between the person groups and item types.

Description of the Saltus Model

The connection between hierarchical theories of development and the realities of data collected as responses to tasks and items is, in the Saltus model, provided by a probabilistic formulation that falls within the family of psychometric models first described by George Rasch (1960/1980). In these models, each person has an ability, θ_i and each item a difficulty, δ_j, measured in logits. The difference between the

ability and the difficulty,

$$\lambda_{ij} = \beta_i - \delta_j$$

determines the probability of success of the person on the item through the logistic function:

$$P(\text{success}) = \frac{\exp \lambda_{ij}}{1 + \exp \lambda_{ij}}$$

In Saltus the same formulation is applied; however, the knowledge that the items were designed to test (at least) two different stages of development is used to make the model more sensitive to the theory. The items are classified into types A and B by this knowledge; A for the earlier stage, B for the later stage. The persons are classified by the meaning that their scores would have if the theory were correct; that is, given L items of type A, a person who scores less than or equal to L should succeed only on items of type A, and so, is classified into group I; and a person who scores more than L should have succeeded on all of type A and some of type B, and so, is classified into group II. With this arrangement, the argument of the logistic function is specified as

$$\lambda_{ij} = \beta_i - \delta_j + \gamma_{hk}$$

where the person and item parameters measure ability within group and difficulty within type, and the Saltus parameter, γ_{hk}, measures the effect on probability of success contributed by membership of person group h for item type k.

Gappiness and rigidity are expressed in the Saltus model as segmentation and asymmetry. Segmentation is the extent to which the item types separate the logit scale into distinct segments, and is indicated by the segmentation index, which is the distance between the most difficult item of type A and the easiest item of type B. If the segmentation index is large and positive, the two item types are clearly separated into distinct regions on the logit scale. If it is zero or negative, the item types occupy the same region on the logit scale. Asymmetry is the relative difference in difficulty of the item types from the perspectives of the two person groups. When the asymmetry index is zero, the Saltus model is equivalent to the simpler Rasch model, which can be interpreted to mean that the difference in difficulty between the two item types is the same for both person groups. When the asymmetry index is positive, the group I persons see the item types as being further apart in difficulty than do the group II persons. This pattern is typical of hierarchical development: the upper stage items are near to impossible for persons at the lower stage, but persons at the upper stage, while finding the upper stage items of medium difficulty, still make a certain amount of 'human error' on the lower stage items. This diminishes the observed difference in difficulty of the item types. This pattern is also manifested in a jump in the predicted probability of

success at the border between the two groups that is not present when the asymmetry index is zero.

A negative asymmetry index, in contrast, indicates that the group I students see the item types as closer together than do the group II students. This is not consistent with rigidity as it implies that some group I students will find some type B items easier than some group II students. This can be caused by a flaw in the item design such as a tendency to elicit guessing. The asymmetry index is the difference between the group I gap, which indicates how hard an average type B item is for an average group I person, and the group II gap, which indicates how easy an average type A item is for an average group II person.

As the item types and person groups have been specified to be consistent with the assumption that a hierarchical theory of development describes the performance of the persons on the items, lack of fit to the model, for either persons or items, indicates some failure in this assumption. Such misfit is not necessarily evidence that the postulated theory is not hierarchical; the problem could lie in the design of the items that are meant to bring the theory to life. Thus, in the search for confirmation of a theory, the Saltus model can contribute not only by providing estimates for a model of person behaviour, but also by providing an indication of the degree to which the data conform to this model.

The Saltus model was estimated with an iterative maximum likelihood algorithm called UCONG that commences with an approximate solution based on PROX. Fit statistics based on the discrepancy between predicted and observed response patterns can be calculated with respect to each person group and item type, allowing the evaluation of the extent of consistency of the data with the estimated model, and providing a framework for diagnosing flaws in items and unusual behaviour in persons. When no group I person is correct on any type B item, or every group II person is correct on all type A items, the Saltus estimation procedure does not work: data sets for which this occurs are called 'intractable'. In such cases, the difference between the two gaps has become infinite (typically, it was found that the BI location became positively infinite). Thus, the step has become 'impossible' for group I students, and the hierarchy has clearly been established. Saltus cannot estimate this infinite gap because the probabilistic assumptions of the model do not hold here. However, as this situation represents what might be called a 'perfect' hierarchy, attention should be given to the construction of an alternate algorithm that will allow the estimation of the non-infinite parameters of this 'perfect' step.

Application of Saltus

The two pieces of educational research which were used to explore the application of Saltus to theories of hierarchical development were chosen because in each case the

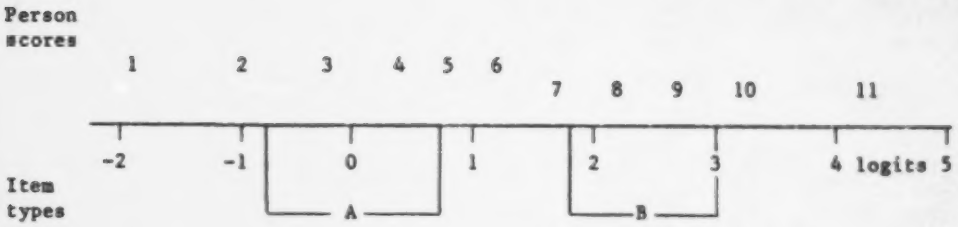


Figure 5.1 Rasch Estimates for 3-digit Subtraction Items

researchers insisted on an explicit and examinable link between the hierarchical theory and the items intended to realize that theory. Research that uses standard published tests of cognitive ability would, in general, not be suitable for Saltus analysis, because standard tests are seldom grounded in explicit theory.

The first data set analysed with Saltus was part of a subtraction sequence assembled in accordance with the learning hierarchy theory of Gagné (Izard et al., 1983). The constructed response items were designed at the Australian Council for Educational Research and administered to third and fourth year students in schools in Victoria and New South Wales. Interest focussed on the transition from being able to solve subtraction problems without regrouping (item type A) to being able to solve problems for which regrouping was needed (item type B); this step was duplicated for subtraction items with both two and three digits. In comparison with the simpler Rasch analysis, three differences were noted. First, the estimates of item difficulty and person ability implied an interpretation that reflects the relationship between the two models. The Rasch analysis logit scale, which is illustrated in Figure 5.1 indicated that both item types were relatively homogeneous in difficulty, that they segmented the logit scale, and that the difference between the means of the two item types was 2.36 logits. These estimates give a probability of success of 0.07 for an average group I person on an average type B item, and a probability of success of 0.94 for an average group II person on an average type A item.

The Saltus analysis logit scale, which is illustrated in Figure 5.2, also indicated that the item types were relatively homogeneous in difficulty, and showed a much stronger segmentation of the logit scale for group I, but a weaker segmentation for group II. The difference between the means of the two types was 4.99 logits for group I and 1.38 logits for group II. The probability of success of an average group I person on an average type B item has become more extreme: 0.01. The probability of success of an average group II student on an average type A item has become less extreme: 0.86. This can be interpreted to mean that, for those who cannot regroup, the regrouping items are almost impossibly hard, and in particular, much harder than the non-regrouping items; but for

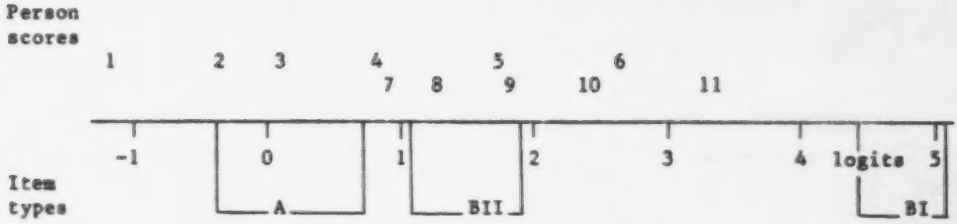


Figure 5.2 Saltus Estimates for 3-digit Subtraction Items

those who can regroup, the difficulty of the non-regrouping items approaches that of the regrouping items, perhaps because of sloppiness, faulty recollection of tables, and other factors commonly labelled 'human error'. Thus, the Rasch model is estimating a model that 'averages' the effects of the two item types.

These differences in pattern might not be meaningful if there were no gain in fit by Saltus over the Rasch model. This leads to the second and third differences between the two. The second difference was that the fit statistics for the Rasch model showed strong negative misfit for all but one of the type B items. This is interpretible as a clue to the existence of rigidity on the step immediately below these items, and this hint was confirmed by the shift of these misfit statistics to unremarkable levels in the Saltus results.

The third difference is in the total log-likelihood for the two analyses: this is an overall measure of fit that takes into account the extent to which every person misfits the model. Twice the difference between the two log-likelihoods provides a likelihood ratio statistic which can be compared to a Chi-Square distribution on one degree of freedom; the obtained value of 37.92 indicates an improvement of fit by the Saltus model which is significant at the 0.001 level.

The second data set analysed by Saltus was produced by three Piagetian tasks modified by R.S. Siegler (1981) to test an adaptation of Piagetian theory called 'rule-assessment'. The rule-assessment model postulates a sequence of four rules which the subjects will use to solve certain types of items, and is tested through an arrangement of apparatus and item types intended to reveal the rule which a subject has attained without the need for further interview. The first task was the prediction of the movement of a balance scale under varying conditions of weights and distances. The transition from Rule I to Rule II was well-segmented and gave a strong positive asymmetry index. The Saltus logit scale gave a pattern of results similar to Siegler's predictions. The probability patterns are shown in the top row of Figure 5.3: the important features are the plateaux in the probability curves for the Rule I items (item type E) for scores 3 to 7, and the jump in the probability curves for the Rule II

items (item type S) near score 4. There were not enough item types to distinguish Rule II from Rule III, but the double step, from Rule II to Rule IV was examined by two Saltus analyses. One gave a negative asymmetry index and the other, although the item types were well segmented and the asymmetry index was positive, gave a wide range of difficulty for the higher problem type. The negative asymmetry index was associated with the CS item types which were designed to elicit guessing from subjects at a certain rule level, so the result matched Siegler's prediction in this respect. The probability curves for this analysis are shown in the middle row of Figure 5.3; the squiggle in the curve for the CS items is caused by the negative asymmetry index. But the prediction for the other item type (S) did not give the hump in probability predicted by Siegler. The lack of homogeneity of the items in the highest step resulted in a pattern of probability, shown in the bottom row of Figure 5.3, that did not match Siegler's predictions for either item type. These results are not conclusive: the first step in the rule-assessment hierarchy matched Siegler's predictions and satisfied the requirements for being a step from a hierarchy. The upper steps matched Siegler's predictions in part. Although they exhibited segmentation, one gave a negative asymmetry index and the other gave a positive asymmetry index that had little influence because of the heterogeneity of the most difficult of the item types.

The same series of analyses was performed on the second task, a problem of predicting the length of a shadow cast by a cross-bar where both the bar and its distance from the light source could be varied. The Saltus results for this task were the same as those for the first task. The third task involved deciding which of two piles composed of red and blue coloured marbles gave a better chance of picking a red marble on a random choice; the variables that were manipulated were the number of red marbles and the number of blue marbles. A different arrangement of problem types was designed for this task, allowing four Rules to be distinguished. The step from Rule I to Rule II was well segmented and showed a strong positive asymmetry index. The linked logit scale for the probability task is illustrated in Figure 5.4. The extra problem type allowed the Rule II to Rule III step to be investigated; but these problem types did not separate the subjects at these Rule levels. The step from Rule III to Rule IV gave a negative asymmetry index indicating that some guessing was occurring, contrary to Siegler's predictions.

Overall, these three tasks show consistency: the Rule I to Rule II step is hierarchical; the existence of the Rule II to Rule III step received no support from the one task for which Saltus could examine the evidence; the Rule III to Rule IV step, though segmented, does not give a hierarchical pattern of results. When the three tasks were linked on a single logit scale, it was found that although the probability task was slightly harder to start but easier to master than the other two, the three tasks are consistent in their placement of the person groups. The pattern revealed suggests that

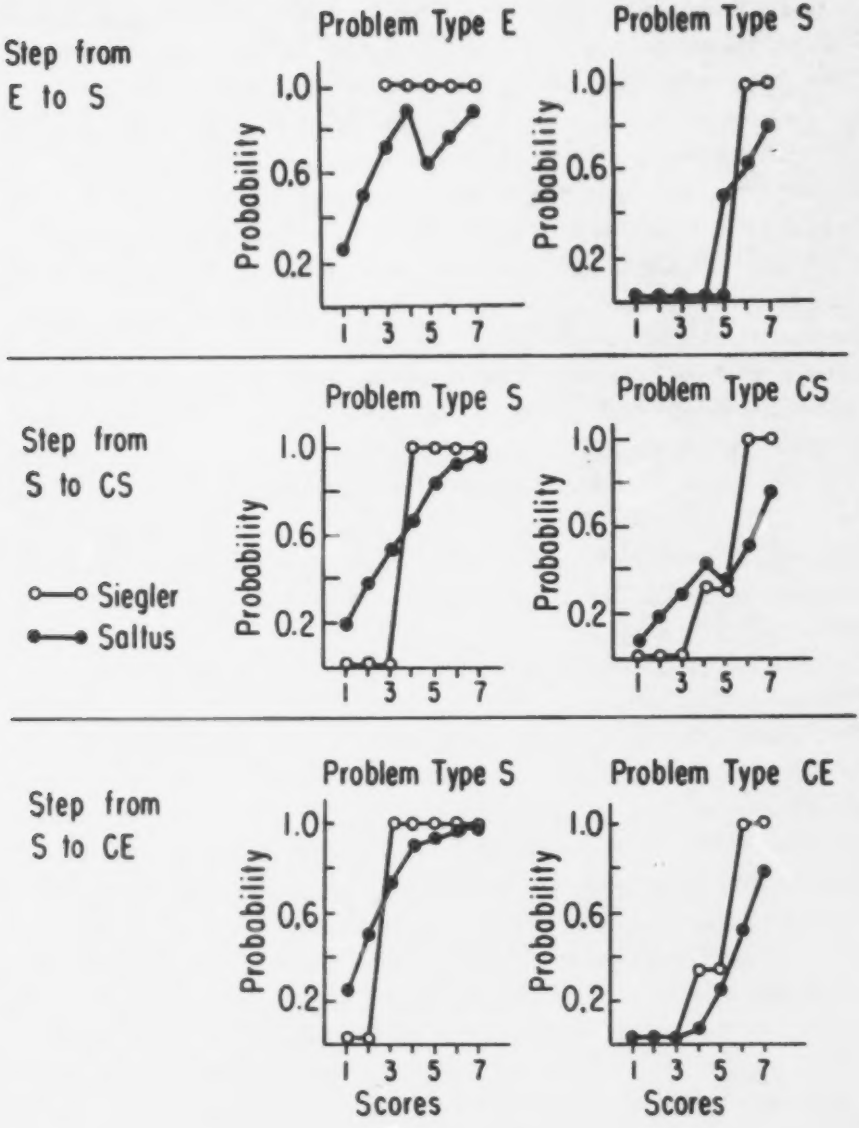


Figure 5.3 Siegler vs. Saltus for Balance Scale

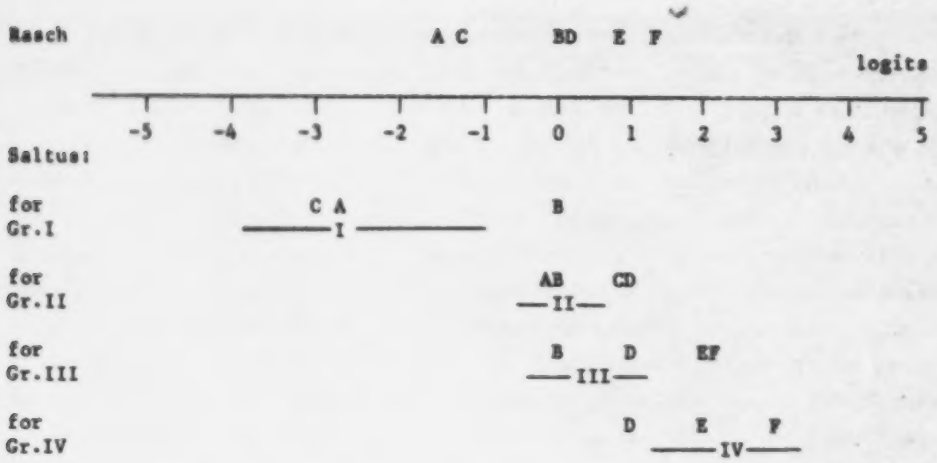


Figure 5.4 Linked Logit Scale for the Probability Task

the original Piagetian classification into three stages (i.e. Rules II and III collapsed) is the more accurate way of representing the rule-assessment data.

In these two analyses, Saltus has demonstrated its ability to respond to the theoretical structures of educational and psychological researchers. For the Gagné subtraction data, a hierarchical step was identified and investigated in a range of contexts. The Siegler rule-assessment data showed that two of the postulated stages could be collapsed to conform to a Piagetian classification. Saltus also demonstrated its relationship to the Rasch model: when asymmetry is zero, the two models give the same results; when asymmetry is strongly positive, the Saltus model gives a pattern of results more complex than the Rasch results, which allows a specifically hierarchical interpretation. The Rasch estimates approximated an average of those for Saltus, but do not give a good a fit as Saltus when asymmetry is strongly positive. The problem of under-estimation of the group I gap when the group II gap is small was also investigated. It was found that a correction could be deduced from a series of tailored simulations when the estimated group I gap was not too large. When the group I gap is large, however, such a solution may not be possible, but the context of the analysis may indicate that any correction would not alter the practical interpretation.

Implications of the Research

There are some avenues for further research along the lines of Saltus that may be fruitful. Once a sound and reliable hierarchy has been established using Saltus, an immediate application would be to the long term monitoring of persons as they passed

through the hierarchy. Thus, the work done by Saltus in identifying the stages of the hierarchy could be applied to the study of change within individuals. A potential adaptation of Saltus is to the situation where there are two indicators of a hierarchy, such as a Piagetian interview and a pencil and paper test. One of the indicators could be used to classify the persons, and this could be used to ascertain the agreement between the two classifications. This would be a useful validation technique for new instruments.

The co-ordination of a psychometric model with psychological and educational theories has not been without cost. The theories did not fit exactly into the form needed by Saltus, and some item designs employed were insufficiently free of guessing to allow a clear-cut interpretation. Whether these are problems with the theories or problems with Saltus depends on your point of view. Interpretation of results is more complex than the simpler Rasch alternative, although Saltus indicates when the simpler model is sufficient. The gains from the application of Saltus have been:

1 the introduction of psychometric ideas into the design of instruments used for the investigation of hierarchies,
2 the development of a meaningful graphical representation of the hierarchy on the logit scale, and
3 the addition of a probabilistic framework for the evaluation and interpretation of persons and items that do no fit the hierarchy.

It is hoped that the presentation of this model has contributed to the value of the pieces of substantive research analysed. It was the high quality of the original work that allowed Saltus to search for patterns of agreement and discrepancy. It is also hoped that this demonstration of an adaptation of the Rasch model will encourage further adaptation of this excellent model to specific measurement and research situations.

REFERENCES

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. Journal of the Royal Statistical Society: Series B, 34, 42-54.

Brainerd, C.J. (1977). Cognitive development and concept learning: an interpretive review. Psychological Bulletin, 84, 919-939.

Brainerd, C.J. (1978). The stage question in cognitive-development theory. Behavioural and Brain Sciences, 1, 173-181.

Bryant, P.E. (1974). Perception and understanding in young children. New York: Basic.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. British Journal of Mathematical and Statistical Psychology, 32, 113-120.

Erikson, E.H. (1956). Childhood and society. New York: Norton.

Fischer, G.H. (1972). Conditional maximum likelihood estimation of item parameters for a linear logistic test model (Res. Bull. No. 9/72). Vienna: University of Vienna, Psychological Institute.

Fischer, G.H. (1973). The linear logistic test model as an instrument of educational research. Acta Psychologica, 37, 359.

Flavell, J.H. (1963). The developmental psychology of Jean Piaget. Princeton, New Jersey: Van Nostrand.

Flavell, J.H. (1971). Stage-related properties of cognitive development. Cognitive Psychology, 2, 421-453.

Flavell, J.H. and Wohwill, J.F. (1969). Formal and functional aspects of cognitive development. In D. Elkind & J.H. Flavell (Eds.), Studies in cognitive development (pp.67-220). New York: Oxford University Press.

Gagne, R.M. (1962a). The acquisition of knowledge. Psychological Review, 69, 355-65.

Gagne, R.M. (1962b). Military training and principles of learning. American Psychologist, 17, 83-91.

Gagne, R.M. (1968). Learning hierarchies. Educational Psychologist, 6, 1-9.

Gagne, R.M. and Bassler, O.C. (1963). Study of retention of some topics of elementary non-metric geometry. Journal of Educational Psychology, 54, 123-31.

Gagne, R.M., Mayor, J.R., Garstens, H.L. & Paradise, N.E. (1962). Factors in acquiring knowledge of a mathematical task. Psychological Monographs, 76 (Whole No. 526).

Gagne, R.M. and Paradise, N.E. (1961). Abilities and learning sets in knowledge acquisition. Psychological Monographs, 75 (Whole No. 518).

Inhelder, B. & Piaget, J. (1959). La genese des structures logiques elementaires: classification et seriation (The early growth of logic in the child: classification and seriation). Neuchatel, France: Delachaux et Niestle.

Izard, J., Farish S., Wilson, M., Ward, G. and Van der Werf, A. (1983). RAPT in subtraction: Manual for administration and interpretation. Melbourne, Australia: Australian Council for Educational Research.

Keating, D.P. (1980). Adolescent thinking. In J.P. Adelson (Ed.), Handbook of Adolescence. New York: Wiley.

Kendall, M.G. and Stuart, A. (1969). Advanced theory of statistics (Vol. 3, 3rd ed.). London: Griffin.

Nagel, E.K. (1957). Determinism and development. In D.B. Harris (Ed.), The concept of development (pp.15-24). Minneapolis, Minnesota: University of Minnesota Press.

Neimark, E.D. (1975). Intellectual development in adolescence. In F.D. Horowitz, Review of child development research (Vol.4, pp.541-594). Chicago: University of Chicago Press.

Oxford University Press. (1961). Oxford English Dictionary. Oxford: Author.

Piaget, J. (1960). The general problems of the psychological development of the child. In J.M. Tanner and B. Inhelder (Eds.), Discussions on Child Development: Vol. 4. The fourth meeting of the World Health Organisation Study Group on the Psychobiological Development of the Child, Geneva, 1956, (pp.3-27). London: Tavistock.

Piaget, J. and Inhelder, B. (1971). Closing remarks. In D.R. Green, M.P. Ford and G.B. Flamer, Measurement and Piaget (pp.210-213). New York: McGraw-Hill.

Pinard, A. and Laurandeau, A. (1969). 'Stage' in Piaget's cognitive-developmental theory: Exegesis of a concept. In D. Elkind and J.H. Flavell (Eds.), Studies in cognitive development (pp.121-170). New York: Oxford University Press.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press. (Original work published 1960).

Siegler, R.S. (1981). Developmental sequences within and between concepts. Monograph of the Society for Research in Child Development, Serial No.189, 46(2).

Spada, H. (1977). Logistic models of learning and thought. In H. Spada and W.F. Kempf (Eds.), Structural models of thinking (pp.227-262). Bern, Switzerland: Hans Huber, 1977.

Spada, H. and Kluwe, R.H. (1980). Two models of intellectual development and their reference to the theory of Piaget. In R.H. Kluwe and H. Spada (Eds.), Developmental models of thinking (pp.1-32). New York: Academic Press.

Stendler, C.B. (1967). Piaget's developmental theory of learning and its implications for instruction in science. In E. Victor and M. Lerner (Eds.), Readings in science education for the elementary school (pp.334-335). New York: Macmillan.

Weber, M. (1949). 'Objectivity' in social science and social policy. In E.A. Shils and H.A. Finch (Trans. and Eds.), Max Weber on the methodology of the social sciences (pp.93-142). Glencoe, Illinois: The Free Press. (Original work published 1904).

White, R.T. (1981). Achievements and directions in research on intellectual skills. The Australian Journal of Education, 25, 224-237.

Wilson, M.R. (1984). A Psychometric Model of Hierarchical Development. Unpublished PhD thesis, University of Chicago, Department of Education.

Wright, B.D. (1982). Person Fit: How and Why. Paper presented at the meeting of the American Educational Research Association meeting, New York.

Wright, B.D. and Douglas, G.A. (1977). Best procedures for sample-free item analysis. Applied Psychological Measurement, 1, 281-294.

Wright, B.D. and Masters, G.N. (1982). Rating scale analysis. Chicago: MESA Press.

Wright, B.D., Masters, G.N. and Ludlow, L.H. (1981). Credit: A Rasch program for ordered categories (Computer program). Chicago: University of Chicago, Department of Education, MESA Psychometric Laboratory.

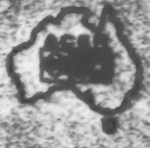
Wright, B.D. and Panchapakesan, N.A. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B.D. and Stone, M.H. (1979). Best test design. Chicago, MESA Press.

Measuring Stages of Growth: A Psychometric Model of Hierarchical Development describes a statistical model for analyzing the kind of data one can expect to find when measuring cognitive development, especially when the development is expected to occur in stages such as those postulated by Piaget.

The model is applied to the learning of subtraction and to three Piagetian tasks; the results indicate that learning has in fact occurred in spurts, but not exactly as predicted by the theory.

ISBN 0 85563 397 2



AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH

BEST COPY AVAILABLE